https://doi.org/10.19195/0080-3626.63.6

MARTA TOMALSKA

TRWAŁOŚĆ ZASOBÓW INTERNETOWYCH NA PRZYKŁADZIE DOSTĘPNOŚCI WITRYN INTERNETOWYCH CYTOWANYCH W „ROCZNIKACH BIBLIOTECZNYCH" W LATACH 1999–2016

Znaczenie dbałości o trwałość zasobów. Sposoby archiwizowania publikacji elektronicznych. Przykłady inicjatyw ratujących od zapomnienia informacje publikowane w sieci. Badanie na przykładzie dostępności witryn internetowych cytowanych w „Rocznikach Bibliotecznych" w latach 1999–2016 (liczba linków, aktywność linków, dostępność linków, typy cytowanych źródeł).

SŁOWA KLUCZOWE: trwałość zasobów internetowych, archiwizacja dokumentów elektronicznych, linki, „Roczniki Biblioteczne"

WSTĘP

Niniejszy artykuł powstał, aby udowodnić potrzebę archiwizacji publikacji funkcjonujących w formie cyfrowej. W tekście poruszono kwestię prawidłowego gromadzenia i udostępniania materiałów, a także opisano przykłady inicjatyw dbających o trwałe zachowanie informacji internetowej. Kolejne fragmenty artykułu prezentują autorskie badanie empiryczne, które porusza problem zależności między datą udostępnienia wybranego dokumentu w Internecie a okresem jego dostępności dla odbiorcy. Opisano w nich zarówno metodę wykorzystaną w przeprowadzonym eksperymencie naukowym, wyniki, jak i ich znaczenie dla ogółu problematyki podejmowanej w tejże pracy.

Współcześnie najpopularniejszą formę informacji stanowi ta przekazywana przez Internet (bez względu na to, czy jest to komunikat *born digital*, czy jedynie zdigitalizowana postać dokumentu pozostającego w formie materialnej). Podczas tworzenia treści o różnorodnym charakterze najistotniejsze pozostaje nie tylko zapewnienie ich jakości, lecz także dostępności dla użytkownika. Wymaga to nie tylko ich odpowiedniego gromadzenia, ale też należytego przechowywania i archiwizowania. Sterowanie procesami zarządzania różnego rodzaju dokumentami jest szcze-

ROCZNIKI BIBLIOTECZNE, R. LXIII: 2019

gólnie ważne z punktu widzenia przekazów publikowanych w Internecie, który (niestety) wciąż uznaje się za nośnik o niskim wskaźniku stabilności i trwałości, co wynika między innymi z szybkiego tempa przemian następujących w jego ramach.

W naszym kraju aktualnie nie są podejmowane żadne kroki mające na celu zgromadzenie i zarchiwizowanie zasobów należących do „polskiego" Internetu (a więc domen z rozszerzeniem .pl). Nieśmiałe działania w tym zakresie podejmowało Narodowe Archiwum Cyfrowe, ale koncentrowały się one na bardzo małym wycinku sieci i nie przyniosły satysfakcjonujących efektów. Niemniej jednak rozwijanie projektów zapewniających użytkownikom stały dostęp do witryn internetowych jest koniecznością, co prezentują między innymi wyniki przeprowadzonego badania, według których aż 63% odsyłaczy traci na swojej aktywności, przez co maleje dostęp do zawartości WWW. Dbałość o prawidłowe archiwizowanie zasobów oraz stosowanie odpowiedniego systemu identyfikowania dokumentów jest zatem obowiązkiem, jeśli chcemy jako społeczność zapewnić sobie ciągłą możliwość korzystania z informacji opublikowanych w Internecie.

ZNACZENIE DBAŁOŚCI O TRWAŁOŚĆ INFORMACJI

Długoterminowa archiwizacja informacji polega na odpowiednim przechowywaniu dokumentów, a także ich ochronie i zabezpieczeniu. W jej ramy ujmowane są wszelkie działania mające na celu zapewnienie możliwości korzystania z określonych publikacji zapisanych w taki sposób, iż są one nieograniczone w czasie swojego trwania lub aktywne do momentu możliwie najbardziej odległego w przyszłości[1]. Współcześnie termin ten bezpośrednio wiąże się z dbałością o trwałość treści przechowywanej przeważnie w postaci cyfrowej. Powodowane jest to głównie przekonaniem, że dane utrwalone w ten właśnie sposób mogą przez lata, o ile będą dostatecznie dobrze chronione, pozostawać dostępne dla wielu użytkowników.

Na tym etapie rozwoju naszej cywilizacji należy przede wszystkim wziąć pod uwagę trwałość nośników, na których cyfrowe dane są zapisywane, a także urządzeń pozwalających na ich wykorzystanie, czyli komputerów, serwerów oraz całej sieci. Są to elementy wymagające nie tylko stałego zasilania elektrycznego, lecz także wykazujące stosunkowo niską odporność na uszkodzenia mechaniczne oraz innego typu awarie.

Technologia ma swoje ograniczenia, związane głównie z dużą dynamiką zachodzących w jej obrębie zmian. Co roku programiści oraz producenci urządzeń elektronicznych wprowadzają na rynek coraz to nowszy sprzęt, a także oprogramowanie. Systemy oraz sprzęt informatyczny ulegają szybkiemu starzeniu się i wychodzą bezpowrotnie z użytku. W ten sposób na przykład dane zapisane

[1] A. Fajfer et al., *Trwała ochrona zasobów cyfrowych — podstawowe pojęcia*, „Biuletyn EBIB" 2014, nr 9 (154), s. 2–3, http://open.ebib.pl/ojs/index.php/ebib/article/view/311/481 [dostęp: 9.11.2018].

na bardzo popularnych kiedyś kartach perforowanych czy dyskietkach są dziś praktycznie nie do odtworzenia. Podobna sytuacja może wystąpić w kontekście CD-ROM-ów, które wychodzą z obiegu, o czym świadczy między innymi fakt, że wiele współczesnych laptopów nie jest wyposażonych w stacje napędowe pozwalające na ich odczytanie. Wszelkiego rodzaju programy ułatwiające korzystanie z poprawnie zakodowanych informacji również są w fazie ciągłych zmian. Sprawa ta nie dotyczy jedynie zastępowania jednego oprogramowania zupełnie nowym, ale też aktualizacji tych samych aplikacji[2].

Większość użytkowników sprzętu elektronicznego przyzwyczaiła się do jego ciągłej wymiany. Stało się to tak powszechną praktyką, że duża część z nas nie sprzeciwia się temu procesowi, nie dbając równocześnie o przenoszenie własnych danych ze starszych nośników na nowsze. Kwestie te są poruszane w dyskursie naukowym między innymi w kontekście osobistej archiwistyki cyfrowej[3]. Jeśli natomiast mówimy o treściach traktowanych jako pewien element dziedzictwa kulturowego, wówczas przypadki utracenia danych nie powinny mieć miejsca.

Aby uniknąć sytuacji, w której tracimy dostęp do zapisanych w postaci cyfrowej dokumentów, możliwe jest zastosowanie kilku rozwiązań. Jednym z nich jest stałe odświeżanie nośnika przechowującego dane, polegające na dokonywaniu wszelkiego rodzaju aktualizacji mających na celu zapewnienie mu jak najdłuższej żywotności. Takie zabiegi okazują się czasem niewystarczające. W tym wypadku należy zmienić generację nośnika, czyli po prostu przenieść potrzebne nam informacje na nowy, lepszy sprzęt. Innym sposobem jest migracja danych, czyli dokonanie ich konwersji do kolejnego formatu. Generuje to jednak koszty, a także problemy związane z koniecznością instalowania nowych oprogramowań mogących nie poradzić sobie z odtworzeniem różnych formatów dokumentów. Rozwiązaniem tej sytuacji wydają się aplikacje emulujące, czyli symulujące inny, przeważnie starszy, program. Jedną z opcji może być też utrzymywanie komputerowych muzeów, w ramach których gromadzone będą eksponaty, stanowiące oryginalne nośniki wraz ze sprzętem i oprogramowaniem potrzebnym do odtworzenia danych[4]. Do takich inicjatyw można zaliczyć między innymi Computer History Museum, DigiBarn Computer Museum czy The National Museum of Computing, działające na terenie Stanów Zjednoczonych[5]. Wymienione alternatywy (szczególnie te ostatnie) są

[2] T. Parkoła, *Długoterminowe przechowywanie cyfrowego dziedzictwa kulturowego*, „Biuletyn EBIB" 2014, nr 9 (154), s. 3, http://open.ebib.pl/ojs/index.php/ebib/article/view/303/477 [dostęp: 9.11.2018].

[3] M. Wilkowski, *Od osobistej archiwistyki cyfrowej do edukacji medialnej*, „Biuletyn EBIB" 2014, nr 6 (151), http://open.ebib.pl/ojs/index.php/ebib/article/view/274/436 [dostęp: 9.11.2018].

[4] A. Januszko-Szakiel, *Archiwistyka cyfrowa. Długoterminowa ochrona dziedzictwa nauki i kultury*, Warszawa 2017, s. 90–91.

[5] M. Góralska, *Imperatyw pamięci i akceptacja zapomnienia w epoce cyfrowej*, [w:] *Nauka o informacji w okresie zmian: informatologia i humanistyka cyfrowa*, red. B. Sosińska-Kalata, M. Przastek-Samokowa, Z. Wiorogórska, Warszawa 2016, s. 56–57.

bardzo kosztochłonne. Nie każdego stać na wymianę sprzętu w bardzo szybkim tempie (choć, jak już zaznaczono, większość z nas przyzwyczaiła się do takiej konieczności), a już na pewno nie na utrzymywanie wcześniejszych sprzętów w formie pozwalającej na ich prawidłowe funkcjonowanie.

Z pojęciem archiwizacji danych bezsprzecznie związany jest termin trwałej ochrony substancji obiektu cyfrowego, definiowany przez Aleksandrę Fajfer, Karolinę Imiołek-Stachurę i innych jako treść publikacji elektronicznej w postaci niezmiennego kodu zerojedynkowego[6]. Zapewnienie dostępu do tego właśnie elementu stanowi kluczowe zadanie w procesie dbałości o informację nie tylko w kontekście archiwizacyjnym, lecz także technologicznym. Pojęcie substancji utworu związane jest również z twierdzeniami Michaela Bhaskara, który w książce *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network* pisał o „wlewaniu" dzieł w różne ramy (wynikające z wykorzystania do odczytu informacji różnych urządzeń)[7]. To, czy dana zawartość będzie możliwa do odtworzenia przez odbiorców, zależy od trwałości nośników zapisu oraz typu wykorzystanych formatów służących do utrwalenia dokumentów, a także zmian w ich zakresie. Współcześnie najpopularniejszą postacią pliku pozwalającą na łatwy dostęp do publikacji cyfrowych jest format PDF. Każdy posiadający komputer z odpowiednim oprogramowaniem i dostępem do sieci może skorzystać z dokumentu zapisanego w ten właśnie sposób, ponieważ może on być odtwarzany przez przeglądarkę internetową. Mimo wszystko często występującym w ramach różnych polskich archiwów cyfrowych formatem jest DjVu, wymagający zainstalowania odpowiedniej wtyczki lub aplikacji.

Koniecznością w wypadku archiwizacji pozostaje więc ciągłe obserwowanie rozwijającego się rynku nowych technologii oraz odpowiednio szybkie reagowanie na zmiany, tak aby informacje mogły jak najdłużej trwać w czasie. Są one bowiem przeważnie dużo trwalsze niż techniki pozwalające na ich przechowywanie i udostępnianie[8].

Substancje publikacji powinny być przede wszystkim użyteczne, autentyczne i integralne. Należy gromadzić informacje mające dla użytkowników jakieś znaczenie. Trzeba zadbać również o to, aby były kompletne. Nie można dopuścić do zmian w dokumentach, co doprowadziłoby do zafałszowania ich treści. Publikacje w formie nadanej im przez autora powinny być dostępne dla jak najszerszego grona odbiorców[9]. Należy także pamiętać o odpowiednim opisie archiwizowanych dokumentów, aby mogły one być z łatwością odszukane przez użytkowników. Znaczącą rolę w tych działaniach odgrywają metadane, które są informacjami

[6] A. Fajfer et al., op. cit., s. 7.

[7] M. Bhaskar, *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network*, London 2016.

[8] A. Fajfer et al., op. cit., s. 7.

[9] M. Nahotko, *Metadane. Sposób na uporządkowanie Internetu*, Kraków 2004, s. 43.

służącymi do opisywania konkretnych publikacji. Taki opis jest kluczowy dla przechowywania, udostępniania, ale też ochrony treści. Współczesny standard w tym zakresie to Dublin Core, zawierający 15 elementów wykorzystywanych podczas charakteryzowania dokumentów[10].

SPOSOBY ARCHIWIZOWANIA ZASOBÓW

Rozwiązaniem problemów związanych z koniecznością zmiany nośnika czy oprogramowania pozwalającego na odtwarzanie konkretnych informacji wydają się magazyny publikacji cyfrowych. W ich ramach można przechowywać, a także udostępniać dokumenty o charakterze elektronicznym.

Powstanie takich systemów jest realizacją rozważań dotyczących gromadzenia, przetwarzania i udostępniania treści za pomocą komputerów, które pojawiły się już w latach 40. XX wieku. Twórcą jednej z takich koncepcji był Edmund Callis Berkeley. Zawarł ją w dziele zatytułowanym *Giant Brains or Machines That Think*[11]. W jego ramach omówił pierwsze komputery, a także ich potencjalne wykorzystanie do stworzenia biblioteki online („elektronicznego mózgu") pozwalającej wyszukiwać poprzez katalog potrzebne nam informacje. Bezpośrednim rozwinięciem tego pomysłu były poglądy Josepha Carla Robnetta Licklidera postulującego konieczność wyciągnięcia treści z książek i innych materialnych utworów oraz utworzenia z tych danych systemów neobibliotecznych.

Kolejnym projektem, który doprowadził do gromadzenia i udostępniania dokumentów w postaci cyfrowej na tak szeroką skalę, był Projekt Gutenberg. Został założony w 1971 roku przez Michaela Harta. Początkowo eksperyment rozwijał się bardzo wolno, co wynikało z ręcznego wprowadzania tekstów do komputera, jednakże współcześnie w ramach tej witryny udostępnianych jest ponad 60 tysięcy zasobów[12]. Podobną inicjatywą był Projekt Runeberg, rozwijany od 1992 roku przez Larsa Aronssona, a także środowisko ochotników.

Następnym etapem ewolucji wcześniej omawianych koncepcji były próby utworzenia narodowych bibliotek bez ścian. Jedną z pierwszych propozycji tego typu był projekt American Memory. Jego pilotażowe programy realizowano już w latach 1990–1994 (zdigitalizowano wówczas różne unikatowe materiały historyczne pochodzące z kolekcji Biblioteki Kongresu, a następnie przekazano je do użytkowania szkołom i bibliotekom w postaci CD-ROM-ów). Natomiast znacz-

[10] *Dublin Core*, Biblioteka Narodowa, https://www.bn.org.pl/dla-bibliotekarzy/normy,--formaty,-standardy/metadane/dublin-core [dostęp: 9.11.2018].

[11] E.C. Berkeley, *Giant Brains or Machines That Think*, New York 1961, https://monoskop.org/images/b/bc/Berkeley_Edmund_Callis_Giant_Brains_or_Machines_That_Think.pdf [dostęp: 9.11.2018].

[12] *Free ebooks — Project Gutenberg*, Projekt Gutenberg, http://www.gutenberg.org/ [dostęp: 9.11.2018].

ny rozwój przedsięwzięcia nastąpił po ogłoszeniu rozpoczęcia National Digital Library Program. Kolejną inicjatywą była Gallica[13], czyli francuska narodowa biblioteka cyfrowa, która została utworzona w roku 1997. Jeśli chodzi o polskie projekty, należy wymienić Polską Bibliotekę Internetową, mającą być realizacją założeń europejskiego programu e-Content. Nie odniosła ona jednak zakładanego sukcesu, w przeciwieństwie do Polony, administrowanej przez Bibliotekę Narodową i zajmującej się od 2006 roku udostępnianiem w postaci cyfrowej dziedzictwa kulturowego Polski[14].

Zanim upowszechniły się profesjonalne cyfrowe biblioteki i repozytoria, podjęto wiele prób cyfrowego archiwizowania oraz sieciowego udostępniania (w kampusach uniwersyteckich) kolekcji elektronicznych wersji artykułów z czasopism naukowych. Wśród nich można wyróżnić: Mercury Electronic Library Project (1989–1992), realizowany przez Carnegie Mellon University, czy The University Licencing Program, rozwijany za pośrednictwem wydawnictwa Elsevier Science Publishers w latach 1991–1995.

Na początku lat 90. XX wieku zaczęły powstawać również otwarte repozytoria dokumentów elektronicznych, określane jako Open Archives[15]. Ich zadaniem było usprawnienie przepływu informacji o badaniach naukowych i ich wynikach. Ta koncepcja ściśle łączy się z ruchem Open Access. Jego głównym celem jest skuteczne rozpowszechnianie treści w Internecie, aby zapewnić trwały oraz powszechny dostęp do publikacji o charakterze badawczym i edukacyjnym. Był to kolejny etap walki o wolność, niezależność oraz swobodny dostęp do informacji, ze szczególnym uwzględnieniem tej naukowej, a także dydaktycznej, ze względu na to, że są one finansowane z funduszy publicznych. Wśród dwóch głównych modeli Open Access można wyróżnić archiwa i repozytoria oraz czasopisma. Zadaniem tych pierwszych jest swobodne udostępnianie zdeponowanych w nich materiałów. W ich ramach nie dokonuje się żadnych działań o charakterze recenzującym. Autorzy mogą dzięki nim informować o postępie swoich prac badawczych sformułowanych w postaci zarówno preprintów, jak i postprintów. W wypadku czasopism stosuje się proces weryfikujący jakość publikowanych w nich artykułów. W niektórych sytuacjach jest również pobierana od autorów opłata za możliwość podzielenia się swoimi dokonaniami[16].

Ze sposobami archiwizowania zasobów wiąże się zatem kilka pojęć. Jak zaznaczono, obiekty mogą być gromadzone w ramach rozmaitych baz danych. Głównym celem tego rodzaju instytucji jest nie tylko przechowywanie treści, ale też dbałość o jej odpowiednie udostępnianie szerokiemu gronu odbiorców. Podstawę ich funk-

[13] Gallica, https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop [dostęp: 9.11.2018].

[14] *O Polonie*, Polona, https://polona.pl/page/o-polonie/ [dostęp: 9.11.2018].

[15] J. Hofmokl et al., *Przewodnik po otwartej nauce*, s. 16, https://repin.pjwstk.edu.pl/files/Przewodnik_po_otwartej_nauce.pdf [dostęp: 23.12.2019].

[16] P. Suber, *A Very Brief Introduction to Open Access*, http://legacy.earlham.edu/~peters/fos/brief.htm [dostęp: 9.11.2018].

cjonowania stanowi sprawne zarządzanie oraz tworzenie systemów pozwalających wyszukiwać i przeglądać publikacje. Gromadzenie oraz przechowywanie dokumentów elektronicznych może się obecnie odbywać w ramach takich form organizacyjnych, jak: archiwum cyfrowe, repozytorium cyfrowe oraz biblioteka cyfrowa.

Archiwum cyfrowe to forma mająca podstawowe znaczenie w kontekście dbałości o trwałość informacji internetowej. To system złożony z ludzi oraz narzędzi, pozwalający na zebranie dokumentów elektronicznych i przeprowadzenie ich przez kolejne etapy rozwoju technologicznego, które zapewniłyby długotrwały dostęp do publikacji. Wśród typów archiwów można wyróżnić: archiwa bibliotek narodowych, archiwa bibliotek uczelnianych, archiwa centrów oraz instytutów badawczych, archiwa sektora administracji, zarządzania i biznesu, archiwa instytucji archiwalnych oraz muzealnych czy archiwa organizowane przez zewnętrznych usługodawców[17].

Kolejną formę, występującą często zamiennie z wyżej opisaną, stanowi repozytorium. Ono również jest traktowane jako technologia umożliwiająca importowanie, eksportowanie, przechowywanie i wyszukiwanie cyfrowych dokumentów. W większości przypadków repozytorium bezpośrednio definiuje się jako repozytorium instytucjonalne gromadzące dorobek konkretnej społeczności naukowej, zrzeszonej na przykład w ramach jakiejś uczelni wyższej. W związku z tym jego głównymi zasobami są przede wszystkim artykuły z czasopism, dysertacje, materiały dydaktyczne, surowe dane z przeprowadzonych badań, sprawozdania oraz raporty[18]. Innym przykładem repozytoriów mogą być repozytoria dziedzinowe, takie jak e-LiS, w ramach których udostępnianie są dokumenty dotyczące konkretnej dyscypliny naukowej.

Publikacje w postaci cyfrowej gromadzi się również w bibliotekach cyfrowych. Ich zasoby są jednak dużo bogatsze, ponieważ tworzą je nie tylko dokumenty powstałe jako obiekty elektroniczne, ale też zdigitalizowane formy dzieł materialnych. Taką działalnością zajmują się przede wszystkim wyspecjalizowane instytucje postępujące w swoich działaniach zgodnie ze standardami bibliotecznymi[19].

Digitalizacja obejmuje kilka różnych etapów. Ich zrealizowanie jest niezbędne do utrwalenia zbiorów na maksymalnie długi czas. Wśród nich przede wszystkim trzeba wymienić sposób dobrania materiałów — digitalizacja może przyjąć charakter selektywny lub masowy. Kolejny punkt to podjęcie decyzji dotyczącej metadanych służących do opisu powstałych obiektów. Koniecznym składnikiem jest też wybór kadry, pomieszczeń i sprzętu digitalizacyjnego. Należy także rozważyć metody oraz sposoby przetwarzania, zapisu i przechowywania danych[20]. Niestety w wypadku ostatniego elementu nie ma jeszcze ogólnie przyjętych standardów

[17] A. Januszko-Szakiel, op. cit., s. 35–36.

[18] *Definicje repozytorium*, Baza Wiedzy Politechniki Warszawskiej, http://repo.bg.pw.edu.pl/index.php/pl/informacje-o-repozytorium-o-rep/definicje-repozytorium [dostęp: 9.11.2018].

[19] A. Januszko-Szakiel, op. cit., s. 33–34.

[20] D. Paradowski, *Digitalizacja piśmiennictwa*, Warszawa 2010, https://www.bn.org.pl/download/document/1342175805.pdf [dostęp: 6.01.2019].

pozwalających zapewnić długotrwałą archiwizację informacji przechowywanych w bibliotekach cyfrowych. Rozwiązaniem tego problemu może być między innymi zapisywanie danych w formatach uniwersalnych i otwartych lub stworzonych specjalnie do realizowania wspomnianego celu (na przykład PDF/A lub ODF). Należy też korzystać z najlepszych nośników służących do przechowywania dokumentów. Pomocne w tym względzie okazuje się uczestniczenie w różnego rodzaju projektach, takich jak Platforma Obsługi Nauki (PLATON) realizowana na podstawie sieci stworzonej przez konsorcjum PIONIER. W jego ramach udostępniana jest usługa pozwalająca bibliotekom cyfrowym zdalnie archiwizować oraz tworzyć zabezpieczenia danych (ang. *backup*)[21]. Innym krokiem może być skorzystanie z oferty Poznańskiego Centrum Superkomputerowo-Sieciowego, a dokładniej Systemu dArceo. Jego głównym celem jest pomaganie instytucjom w zakresie formatów wykorzystywanych do utrwalania zasobów, jak również ich bezstratnej migracji do nowych postaci[22]. Projekt ten powstaje przy współpracy z Open Preservation Foundation, organizacji wspierającej otwarte przechowywanie danych cyfrowych. Do pokrewnych instytucji międzynarodowych należą: IMPACT Center of Competence in Digitisation oraz Digital Preservation Coalition[23].

PRZYKŁADY INICJATYW RATUJĄCYCH ZASOBY OD ZAPOMNIENIA

Biblioteki cyfrowe, repozytoria oraz archiwa w głównej mierze skupiają się na gromadzeniu, przechowywaniu i udostępnianiu plików w postaci cyfrowej. Są to przeważnie kopie obiektów materialnych, takich jak artykuły, czasopisma, książki czy nawet obrazy. Nie należy mieć wątpliwości, iż tego rodzaju inicjatywy są bardzo istotne z punktu widzenia archiwizowania informacji. Jednakże w ich ramach nie gromadzi się zazwyczaj zasobów internetowych nienależących do wspomnianych kategorii. Zdarza się zatem, że treści poszczególnych stron WWW w związku z dużą dynamiką ich zmian, ciągłym procesem aktualizacji, przenoszeniem określonych fragmentów pomiędzy różnymi częściami witryny, a także ich usuwaniem po prostu znikają z sieci. Niemniej jednak wciąż podejmowane są różnego rodzaju działania, które mają na celu uratowanie od zapomnienia informacji składających się na cyberprzestrzeń oraz liczonych w zettabajtach. Ze względu na duże rozmiary przestrzeni internetowej jej archiwizacja w całości wydaje się niemożliwa. Jednak istnieją i wciąż powstają nowe organizacje zajmujące

[21] *Usługi powszechnej archiwizacji*, PLATON, http://www.platon.pionier.net.pl/online/archiwizacja.php [dostęp: 9.03.2019].

[22] D. Witczak, K. Sobkowiak, *Problemy przechowywania danych cyfrowych w bibliotekach*, „Elektroniczne Czasopismo Biblioteki Głównej Uniwersytetu Pedagogicznego w Krakowie" 2014, nr 5, s. 3, 6, http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-e270bf28-b322-47fb-b4a9-8a59456498c1/c/BiE_nr_5_2014_A5_w2.pdf [dostęp: 6.01.2019].

[23] T. Parkoła, op. cit., s. 4–9.

się omawianą kwestią. Ich zadaniem jest dbałość o archiwizowanie informacji stanowiących zawartość rozmaitych portali.

Dynamizm zmian zachodzących w sieci jest ogromny. Zauważają to między innymi tacy badacze jak Maria Anna Jankowska, twierdząca, iż 44% witryn internetowych znika w ciągu jednego roku[24]. Narodziła się zatem potrzeba dbałości o trwałość zasobów cyfrowych. Jej konsekwencją było utworzenie archiwów Internetu. Według Wikipedii istnieje prawie 100 inicjatyw tego typu[25]. Niestety dotychczas nie powstał jeszcze projekt pozwalający na zabezpieczanie całości informacji upublicznianych w Internecie, a także jednoczesne udostępnianie ich użytkownikom w niezmienionej i nieograniczonej wersji. Trzeba jednak w tym miejscu przytoczyć kilka koncepcji będących najpopularniejszymi projektami związanymi z archiwizacją sieci.

Najbardziej znanym przedsięwzięciem w omawianym zakresie jest Internet Archive[26]. Zostało ono założone w 1996 roku przez Brewstera Kahla. To projekt, który powstał jako rodzaj biblioteki internetowej mającej zapewnić dostęp wszystkim zainteresowanym (badaczom, historykom, uczonym, osobom niepełnosprawnym i szeroko pojętej publiczności) do historycznych kolekcji funkcjonujących w formacie cyfrowym[27]. Jednym z najważniejszych elementów omawianego pomysłu jest serwis Wayback Machine. To narzędzie umożliwiające „przeniesienie się w czasie" oraz odczytanie dawniejszych wersji stron internetowych, pod warunkiem że zostały one odpowiednio zaindeksowane. Proces ten wspierany jest przez specjalnie stworzone na potrzeby projektu oprogramowanie typu *open source*[28]. Większość serwisów jest automatycznie zapisywanych przez Wayback Machine. Właściciele portali mogą też zwrócić się do twórców z prośbą o ich zarchiwizowanie. Utrwalone wersje stron internetowych są przechowywane na serwerach zlokalizowanych w trzech miastach Kalifornii. Ich odbicie znajduje się w Egipcie w Bibliotheca Alexandrina. Aby wyszukać interesującą witrynę, należy znać jej dokładny adres URL. Następnie, jeśli stronę zapisano, pojawia się kalen-

[24] M.A. Jankowska, *Biblioteki akademickie — trendy dotyczące zasobów elektronicznych*, 2008, s. 168, http://www.library.put.poznan.pl/konf_idn/art/4_3.pdf [dostęp: 24.11.2018]. Warto zaznaczyć, iż autorka swoje badania opublikowała w 2008 roku. Należy zatem przypuszczać, że żywotność współczesnych witryn internetowych jest znacznie krótsza niż wówczas, ponieważ powstaje ich zdecydowanie więcej, a treści zamieszczane na poszczególnych podstronach dużo częściej ulegają migracjom.

[25] *List of Web archiving initiatives*, Wikipedia, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 24.11.2018].

[26] Witryna projektu dostępna pod adresem: https://archive.org/ [dostęp: 24.11.2018].

[27] *About the Internet Archive*, Internet Archive, http://www.archive.org/about/about.php [dostęp: 24.11.2018].

[28] K. Gmerek, *Archiwa internetowe po obu stronach Atlantyku — Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1 (128), s. 3, http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf [dostęp: 24.11.2018].

darz z oznaczonymi datami, w których została zarejestrowana. Częstotliwość oraz głębokość utrwalania zasobów sieciowych jest bardzo różna.

Wokół tego narzędzia pojawiło się kilka kontrowersji. Większość z nich dotyczyła niechęci twórców stron internetowych do utrwalania publikowanych przez nich informacji. Prośby o usunięcie archiwalnej kopii witryny są respektowane, a uniknąć indeksowania można dzięki umieszczeniu na serwerze pliku „no robots". Problemem, z jakim muszą mierzyć się osoby odpowiedzialne za kreowanie tego projektu, jest nieustannie zwiększająca się liczba stron internetowych, a także tempo ich aktualizacji. W związku z tym, że w ramach przedsięwzięcia, jakimi są Internet Archive i Wayback Machine, gromadzone są całe witryny — wraz ze skryptami, aplikacjami czy interaktywnymi linkami — mogą wystąpić trudności związane z miejscem potrzebnym na przechowywanie tak ogromnych zasobów, jak również funduszami niezbędnymi do ich utrzymania[29]. Opłaca się je jednak pozyskiwać, ponieważ zebrane przez B. Kahla i jego współpracowników dane stanowią wierne odzwierciedlenie Internetu w różnych fazach jego trwania.

Innym przedsięwzięciem jest PANDORA[30], powstająca od 1996 roku w Australii. Za jej funkcjonowanie odpowiadała Australijska Biblioteka Narodowa. Od 2019 roku projekt ten stał się częścią Trove. W jego ramach zbierane są dokumenty dotyczące tego państwa oraz tworzone na jego terytorium. Kluczowym kryterium podczas doboru obiektów jest ich potencjalna przydatność do prowadzenia działalności badawczej. PANDORA jest więc archiwum selektywnym. W związku z brakiem odpowiednich regulacji prawnych dotyczących egzemplarza obowiązkowego w tym zakresie na terenie Australii występuje konieczność pytania właścicieli zasobów o pozwolenie na ich utrwalenie[31]. Jeśli chodzi o przeszukiwanie zbiorów, odbywa się ono poprzez przeglądanie tematyczne lub śledzenie alfabetycznej listy tytułów.

Długotrwała archiwizacja plików cyfrowych jest przedmiotem rozważań i działań również w Czechach. Od 2001 roku tworzy się tam projekt Webarchiv[32]. Obecnie jest on rozwijany przez Czeską Bibliotekę Narodową, chociaż wcześniej w prace zaangażowana była Biblioteka Morawska oraz jeden z instytutów Uniwersytetu Masaryka. Trzeba też dodać, że inicjatywa ta ma wsparcie Ministerstwa Kultury Czech[33].

---

[29] Ibidem, s. 5.

[30] Witryna projektu dostępna pod adresem: https://pandora.nla.gov.au/ [dostęp: 24.11.2018].

[31] L. Derfert-Wolf, *Archiwizacja Internetu — wprowadzenie i przegląd wybranych inicjatyw*, „Biuletyn EBIB" 2012, nr 1 (128), s. 12, http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf [dostęp: 24.11.2018].

[32] Witryna projektu powinna być dostępna pod adresem: https://www.webarchiv.cz/.

[33] G. Gmiterek, *Długoterminowa archiwizacja zasobów cyfrowych*, s. 219, http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.ojs-doi-10_17951_rh_2013_35_213/c/1157-954.pdf [dostęp: 24.11.2018].

Hrvatski arhiv weba[34] to projekt rozwijany przez Narodową i Uniwersytecką Bibliotekę Chorwacką. Został uruchomiony w 2003 roku. Jego celem jest uwiecznienie szczególnie ważnych treści internetowych, takich jak: czasopisma, książki, portale, witryny instytucji, ważnych osób, niektóre blogi itd. W tym kraju strony WWW podlegają bowiem ustawie o egzemplarzu obowiązkowym. Nie jest zatem konieczne ubieganie się o zgodę ich twórców podczas prowadzenia procesu archiwizacji. Znalezienie żądanych dokumentów może się odbyć poprzez wpisanie odpowiedniego tytułu, słowa kluczowego czy adresu URL. Jedną z opcji jest także zaznajomienie się z treściami witryn z wykorzystaniem spisu tytułów lub dziedzin[35].

Kolejnym przykładem projektu, którego główne zadanie to archiwizowanie zasobów Internetu, jest UK Web Archive (UKWA)[36]. Inicjatywę rozwija się od 2004 roku[37]. Głównym robotem pozwalającym gromadzić witryny w jej ramach jest Web Curator Tool. Bazuje ono na wolnym oprogramowaniu Heritrix, stworzonym w Internet Archive[38]. W odróżnieniu od wspomnianej inicjatywy jej brytyjski odpowiednik nie zajmuje się utrwalaniem wszystkich stron internetowych, co czyni go archiwum selektywnym. Wśród udostępnianych kolekcji możemy znaleźć jedynie te, które powstały na terenie Zjednoczonego Królestwa Wielkiej Brytanii i Irlandii Północnej. Materiały powinny być też przydatne do badań naukowych oraz prezentować innowacyjność stron internetowych[39].

Następnym elementem odróżniającym UKWA od Internet Archive, a dokładniej Wayback Machine, jest to, że wszelkie zasoby są zapisywane po uprzednim uzyskaniu zgody twórców. W brytyjskiej inicjatywie inaczej wygląda też kwestia przeszukiwania danych, która jest dość rozbudowana. Wybraną przez nas stronę możemy odszukać nie tylko z wykorzystaniem jej dokładnego adresu URL, lecz także poprzez wyrażenia odwołujące się do tematu czy kolekcji. Możliwe jest zwykłe przeglądanie wspomnianych zbiorów, jak również treści identyfikowanych z konkretnymi dziedzinami czy indeksami.

Archiwizacją stron internetowych w Portugalii zajmuje się od 2008 roku Foundation for National Scientific Computing. Organizacja ta przejęła koncepcję realizowaną od 2001 roku przez Uniwersytet Lizboński. Wspomniany pomysł zakładał utrwalenie wszystkich witryn o rozszerzeniu .pt (właściwym dla tego kraju)

[34] Witryna projektu dostępna pod adresem: http://haw.nsk.hr/ [dostęp: 24.11.2018].

[35] L. Derfert-Wolf, op. cit., s. 14.

[36] Witryna projektu dostępna pod adresem: https://www.webarchive.org.uk/en/ukwa/index [dostęp: 24.11.2018].

[37] W sieci podawane są różne daty założenia, przeważnie jest to rok 2005, który jest początkiem funkcjonowania UK Web Archive Consortium. Natomiast prace nad projektem UKWA rozpoczęły się już w 2004 roku, czego można się dowiedzieć ze strony Biblioteki Brytyjskiej: *UK Web Archive*, British Library, https://www.bl.uk/collection-guides/uk-web-archive [dostęp: 25.11.2018].

[38] K. Gmerek, op. cit., s. 8.

[39] Ibidem, s. 7.

oraz funkcjonujących w innej domenie, ale traktujących o Portugalii. Portuguese Web Archive[40] (obecnie Arquivo.pt) działa z wykorzystaniem oprogramowania Heritrix. Zasoby mogą być przeszukiwane przez słowa kluczowe, adresy URL, format, datę lub domenę[41].

Działania archiwizacyjne podejmują również sami internauci. W Stanach Zjednoczonych rozwija się Archive Team, istniejący od 2009 roku. To przedsięwzięcie zostało utworzone przez Jasona Scotta, który wraz z innymi użytkownikami sieci chciał promować zabezpieczanie dziedzictwa cyfrowego. Ich wysiłek koncentrował się na ochronie treści likwidowanych przez pewne firmy usuwające swoje usługi internetowe[42].

W tym miejscu należy opisać organizacje, które nie zajmują się bezpośrednio archiwizowaniem dokumentów cyfrowych, ale mają za zadanie wspierać ten proces. Wśród nich można znaleźć na przykład Research Library Group. Jednostka ta została założona w 1974 roku przez zespolenie sił bibliotek Uniwersytetu Yale, Uniwersytetu Columbii, Harvardu oraz biblioteki publicznej Nowego Jorku. W 2006 roku połączono ją z Online Computer Library Center. Ta placówka działała już od 1967 roku, początkowo jako Ohio College Library Center[43]. Scalenie tych dwóch podmiotów miało ustabilizować amerykańską politykę dbałości o odpowiednie gromadzenie, udostępnianie oraz archiwizowanie zasobów w postaci cyfrowej.

Ochroną dokumentów sieciowych w Wielkiej Brytanii zajmuje się The Digital Preservation Coalition oraz Digital Curation Center. Obie instytucje powołano do życia na początku XXI wieku. Ich misja skupia się na zabezpieczaniu danych w postaci cyfrowej, wspieraniu instytucji, które zajmują się taką formą informacji, oraz współpracy z innymi podmiotami współtworzącymi współczesny świat ochrony treści elektronicznych[44].

Nestor to niemieckie stowarzyszenie siedmiu bibliotek kooperujących w zakresie ochrony dziedzictwa cyfrowego od 2003 roku. Jego głównym celem jest współdziałanie z innymi instytucjami w wymiarze międzynarodowym. Podobną rolę odgrywa australijska grupa Preserving Access to Digital Information. Jej priorytet to merytoryczne wspieranie jednostek zajmujących się udostępnianiem zasobów sieciowych[45].

W przeciwieństwie do wymienionych państw w Polsce, jak już wspomniano, obecnie nie ma projektu, który zajmowałby się koniecznym archiwizowaniem zasobów polskiego Internetu. Jedynym działaniem w tym zakresie było przeprowadzenie w latach 2009–2010 przez Narodowe Archiwum Cyfrowe programu mającego na celu zarchiwizowanie witryn portali działających jako domeny rządowe, a więc

[40] Witryna projektu dostępna pod adresem: https://arquivo.pt/?l=en [dostęp: 24.11.2018].
[41] L. Derfert-Wolf, op. cit., s. 13.
[42] M. Góralska, op. cit., s. 58.
[43] A. Januszko-Szakiel, op. cit., s. 140.
[44] Ibidem, s. 141.
[45] Ibidem.

z rozszerzeniem .gov.pl. Niestety była to jednorazowa akcja, która nie doczekała się swojej kontynuacji[46]. Jedyną polską stroną WWW pozwalającą zobaczyć archiwalne wersje witryn internetowych jest HistoriaStron.pl[47]. Szkoda, że ta wyszukiwarka w swoim działaniu korzysta nie z własnych zasobów, ale tych zgromadzonych w ramach Google Cache lub Internet Archive. Nie da się jednak ukryć, że opisywane inicjatywy są bardzo potrzebne, aby nie doszło do sytuacji, w której bezpowrotnie utracimy istotne materiały zawarte w polskich zasobach sieciowych.

BADANIA WŁASNE — METODA I WYNIKI

O tym, jak istotny jest to problem, świadczą opisane w niniejszym artykule badania dotyczące dostępności witryn internetowych cytowanych w „Rocznikach Bibliotecznych" w latach 1999–2016. Przeanalizowano częstotliwość przywoływania publikacji o charakterze cyfrowym, aktywność odnośników do nich odsyłających, a także dostępność samych zasobów internetowych. Tym samym starano się wykazać, jak istotne jest prawidłowe archiwizowanie zawartości sieci oraz dbałość o odpowiednie identyfikowanie dokumentów elektronicznych.

„Roczniki Biblioteczne" są periodykiem o długiej tradycji oraz wysokich standardach publikowania. Ukazują się raz w roku i mają ogólnopolski zasięg. Treść poszczególnych zeszytów stanowią przede wszystkim rozprawy z zakresu bibliologii i informatologii[48], ze specjalnym uwzględnieniem bibliografii, nauki o książce oraz bibliotekoznawstwa. Główny trzon publikacji obejmuje oryginalne prace badawcze.

Podczas wyszukiwania w czasopiśmie odsyłaczy do stron WWW skupiono się na przypisach bibliograficznych. Ten rodzaj aparatu naukowego jest zamieszczany najczęściej, po to aby udokumentować cytat, odesłać do literatury źródłowej lub innych prac uwzględniających informacje dotyczące podobnej tematyki. Przypisy sporządzone w stylu tradycyjnym (oksfordzkim) zawierają opisy bibliograficzne artykułów, książek czy też innych dzieł występujących w postaci drukowanej, a coraz częściej także elektronicznej.

Same hiperłącza natomiast, nazywane w polskich opracowaniach najczęściej odnośnikami lub odsyłaczami, mają kierować odbiorcę do dokumentu elektronicznego zamieszczonego w sieci. Najczęściej przyjmują one formę adresu URL, czyli ujednoliconego sposobu prezentowania lokalizacji zasobów dostępnych

[46] S. Zdziebłowski, *Ekspert: potrzeba archiwizowania zasobów internetowych to realny problem*, Nauka w Polsce, 11.01.2018, http://naukawpolsce.pap.pl/aktualnosci/news%2C27860%2Cekspert-potrzeba-archiwizowania-zasobow-internetowych-realny-problem.html [dostęp: 26.11.2018].

[47] Witryna dostępna pod adresem: http://historiastron.pl/ [dostęp: 26.11.2018].

[48] Zgodnie z najnowszym rozporządzeniem Ministerstwa Nauki i Szkolnictwa Wyższego tematy z zakresu bibliologii i informatologii są obecnie częścią nauki o komunikacji społecznej i mediach.

w Internecie[49]. Ich celem jest zarówno wskazanie lokalizacji, jak i umożliwienie użytkownikowi przejścia do witryny, portalu, blogu, artykułu itp., na które powołuje się autor danego tekstu.

Na potrzeby przeprowadzonego badania przeanalizowano numery opublikowane w latach 1990–2016. W doborze materiału źródłowego sugerowano się głównie prawdopodobieństwem wystąpienia w tych wydaniach odwołań do zasobów publikowanych w Internecie. Powodem przyjętych ram czasowych była znajomość daty powstania usługi internetowej World Wide Web. Wspomniany przedział został jednak zawężony do lat 1999–2016, ponieważ w tomach za lata 1990–1998 nie było żadnych odesłań do stron internetowych. W związku z tym, że jedynie numery za lata 2008–2018 są opublikowane bezpłatnie w formie cyfrowej[50], większość materiałów zebrano w sposób tradycyjny poprzez przepisywanie adresów URL z publikacji drukowanych do pamięci komputera, a dokładniej do bazy danych stworzonej z wykorzystaniem programu MS Excel. Każdy link w bazie opatrzono informacjami pozwalającymi zidentyfikować artykuł, z którego pochodzi (zeszyt, tytuł artykułu, autor artykułu i numeru strony). Uwzględniono w niej także wyniki testów — stwierdzenie, czy dany odsyłacz pozostaje aktywny. Uzyskano je za pomocą algorytmu stworzonego we wspomnianym programie przez zakładki *Deweloper* i *Visual Basics*. Jego kod wygląda następująco:

```
Function URLExists(url As String) As Boolean
    Dim Request As Object
    Dim ff As Integer
    Dim rc As Variant

    On Error GoTo EndNow
    Set Request = CreateObject("WinHttp.WinHttpRequest.5.1")

    With Request
      .Open "GET", url, False
      .Send
      rc = .StatusText
    End With
    Set Request = Nothing

    If rc = "OK" Then URLExists = True
    Else: URLExists = False
    End If


    Exit Function
EndNow:
End Function
```

Il. 1. Kod pozwalający programowi MS Excel na sprawdzenie, czy dany adres URL działa

Źródło: opracowanie własne.

[49] T. Berners-Lee, L. Masinter, M. McCahill, *Uniform Resource Locators (URL)*, grudzień 1994, https://tools.ietf.org/html/rfc1738 [dostęp: 16.02.2018].

[50] Dostępne pod adresem: http://rbibl.wuwr.pl/.

Algorytm pozwolił automatycznie otworzyć link w przeglądarce. Jeśli dana strona WWW załadowała się prawidłowo, program uznawał, że odnośnik działa. Jeżeli witryna nie wczytała się poprawnie (pusta strona lub przekierowanie na inną), oznaczało to, że odsyłacz nie jest aktywny, nie prowadzi bowiem bezpośrednio do żadnego konkretnego miejsca w sieci.

Jeśli link został przez program określony jako niedostępny, wówczas następowało ręczne wyszukiwanie zasobów, do których miał odsyłać, z wykorzystaniem przeglądarki Google. Pozwoliło to sprawdzić, czy nadal można dotrzeć do żądanych treści. Tym samym zweryfikowano ich dostępność w Internecie. Jeśli pojawiła się możliwość odnalezienia potrzebnych tekstów lub witryn, zostały one uwzględnione w badaniu i opatrzone stosownym komentarzem.

W tomach objętych badaniami wystąpiły 372 odnośniki[51]. W wyniku testów ustalono, że 196 odsyłaczy wciąż jest aktywnych, natomiast 176 nie działa. Wygenerowane informacje zostały starannie skontrolowane. Sprawdzono również, czy można odnaleźć treści, do których miały odsyłać odnośniki oznaczone przez algorytm jako nieaktywne.

Weryfikacja rezultatów doprowadziła do obniżenia liczby faktycznie niedostępnych zasobów. Osiągnięty wcześniej efekt był prawdopodobnie spowodowany zadaniem zbyt krótkiego przedziału czasowego, po którym aplikacja decydowała, czy dana witryna działa, czy też nie. Spośród wskazywanych przez autorów publikacji elektronicznych dostępnych jest obecnie 235. Na swojej aktywności straciło 137 linków do zasobów Internetu — witryny je zawierające już nie działają lub nie uwzględniają cytowanych treści. Po dodaniu do otrzymanych liczb danych o dokumentach, które wyszukano manualnie za pomocą wyszukiwarki Google, wartości te obniżyły się po raz kolejny. Końcowe rezultaty są następujące: dostęp można uzyskać do 310 zasobów cytowanych przez autorów artykułów z omawianego czasopisma. Nie udało się odszukać 62 cytowanych zasobów. Otrzymane wyniki prezentuje tabela.

Tab. 1. Podsumowanie wyników badań

Liczba linków cytowanych (ogółem)	Liczba linków aktywnych	Liczba linków nieaktywnych	Liczba dostępnych zasobów	Liczba niedostępnych zasobów
372	196 (235)*	176 (137)*	310	62

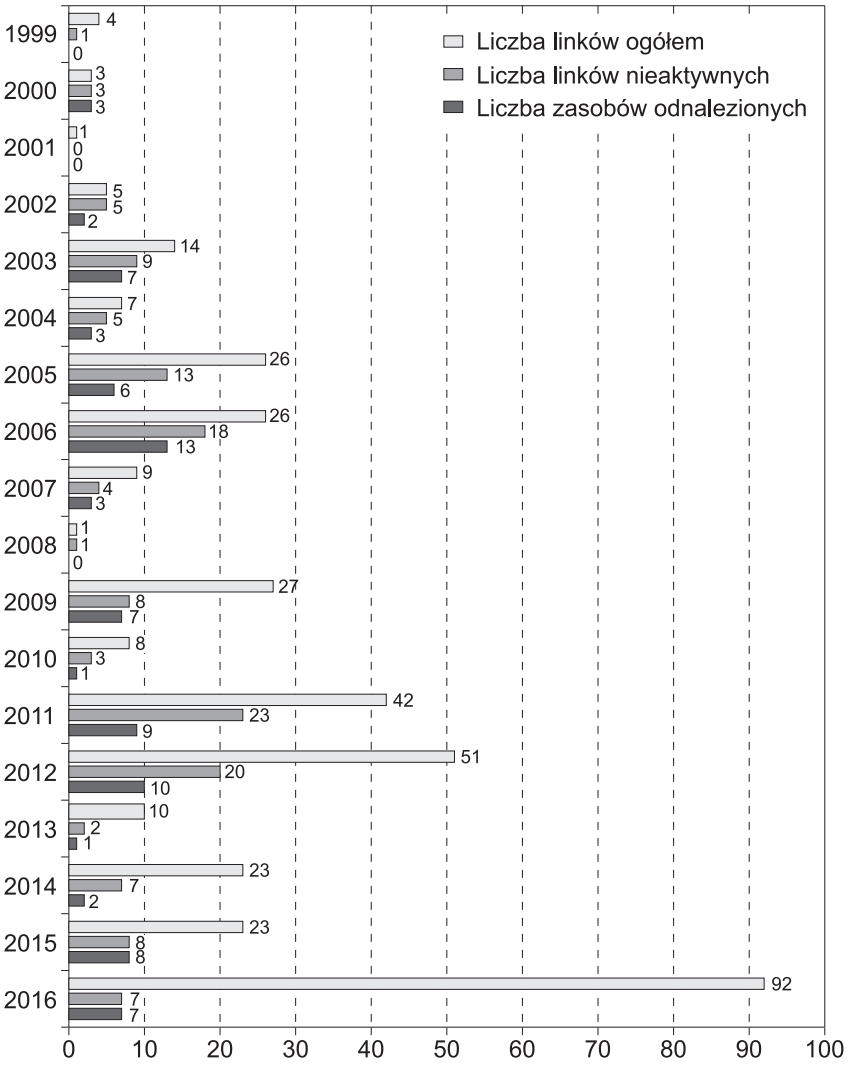
* W nawiasie podano rzeczywistą liczbę linków po zweryfikowaniu ich dostępności.

Źródło: opracowanie własne.

LICZBA LINKÓW

Liczbę odnośników w podziale na aktywne i nieaktywne w poszczególnych tomach „Roczników Bibliotecznych” prezentuje wykres 1.

[51] Prezentowane wyniki dotyczą stanu zakończenia badań (13 maja 2018 roku).



Wykr. 1. Linki w „Rocznikach Bibliotecznych" za lata 1999–2016

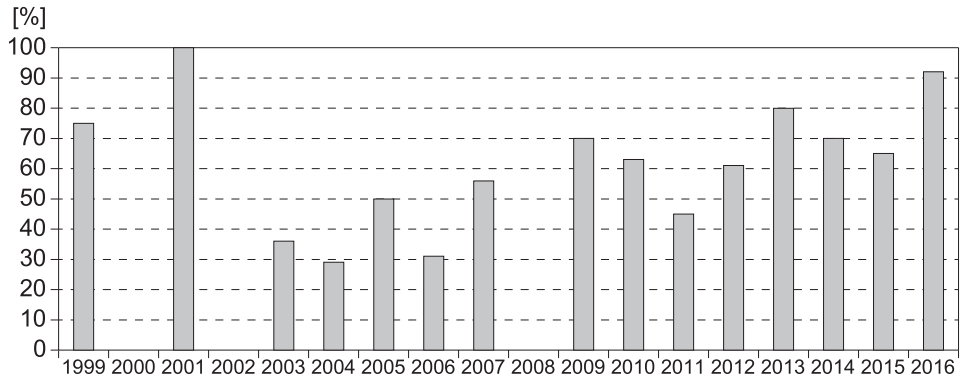
Źródło: opracowanie własne.

Liczba cytowanych dokumentów cyfrowych w poszczególnych numerach analizowanego czasopisma zmieniała się w czasie. Niemniej jednak nie można dostrzec proporcjonalnego do upływu czasu wzrostu odwołań do publikacji w postaci elektronicznej. Rysuje się wyraźna tendencja wzrostowa, co może potwierdzać tezę, zgodnie z którą wraz ze wzrostem popularności Internet staje się bardziej doceniany jako źródło wartościowych materiałów.

Na wykresie można jednak dostrzec kilka „załamań". Jednym z powodów niestałego przyrostu odnośników do stron WWW jest tematyka poszczególnych wydań analizowanego periodyku. Numery, które treściowo są bardziej powiązane z tematami historycznymi, zawierają mniej odnośników do zasobów cyfrowych. Przykładem może być wydanie „Roczników Bibliotecznych" opublikowanych za rok 2008, w którym można znaleźć w przypisach bibliograficznych tylko jeden link. W zeszycie tym opublikowano artykuły poświęcone na przykład znakowi firmowemu Szymona Kempiniego[52] — drukarza krakowskiego działającego w I połowie XVII wieku — lub popularyzacji czytelnictwa wśród chłopów w Królestwie Polskim w deklaracjach, programie i działalności obozu narodowego w latach 1886–1905[53], a więc treści o charakterze historycznym. Podobny przypadek stanowią „Roczniki Biblioteczne" z 2013 roku.

AKTYWNOŚĆ LINKÓW

Kolejnym badanym aspektem była aktywność linków. Na wykresie 2 zaprezentowano stopnień rzeczywistej żywotności adresów URL lokalizujących treści, na które powoływali się autorzy tekstów w „Rocznikach Bibliotecznych".



Wykr. 2. Aktywność linków w „Rocznikach Bibliotecznych" za lata 1999–2016

Źródło: opracowanie własne.

Podczas przeprowadzania badania założono, że im nowszy numer czasopisma, tym większe prawdopodobieństwo, iż cytowane w nim źródła pozostają wciąż aktywne. Przewidywania te jednak się nie sprawdziły. Poczyniono w tym względzie następujące obserwacje:

[52] K. Krzak-Weiss, *Jeszcze jeden polski sygnet o emblematycznych koneksjach (czyli kilka uwag o znaku firmowym Szymona Kempiniego)*, „Roczniki Biblioteczne" 52, 2008, s. 3–13.

[53] A. Karczewska, *„Wszystko dla ludu przez lud". Popularyzacja czytelnictwa wśród chłopów w Królestwie Polskim w deklaracjach, programie i działalności obozu narodowego w latach 1886–1905*, „Roczniki Biblioteczne" 52, 2008, s. 31–45.

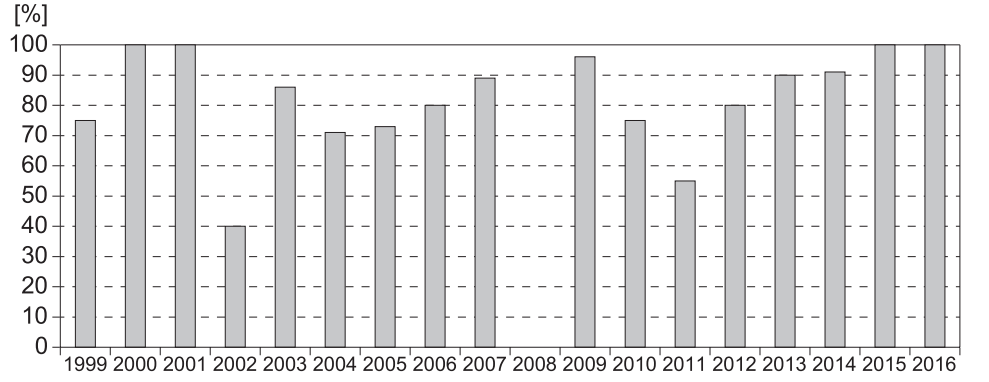
— procent aktywności stron WWW cytowanych w „Rocznikach Bibliotecznych" opublikowanych za lata 1999–2016 ma zmienny charakter. Można dostrzec tendencję wzrostową, ale jest ona poprzecinana licznymi „załamaniami". Najczęściej mamy do czynienia z dostępnością na dużym lub przyzwoitym poziomie, po której następuje krach — bardzo mało stron wciąż działa, a następnie poziom żywotności rośnie itd.;

— żywotność zasobów cyfrowych ma bardzo nierówny poziom — nie można mieć pewności, że jeśli artykuł został zacytowany niedawno, to na pewno jest dostępny. Ma to związek z dużą dynamicznością Internetu i funkcjonujących w jego ramach treści, które często są aktualizowane, archiwizowane, usuwane, przenoszone między poszczególnymi podstronami portali lub witrynami w ogóle;

— niektóre witryny cytowane w początkowych numerach czasopisma wciąż funkcjonują, ponieważ wtedy często odwoływano się do stron głównych, a nie poszczególnych zakładek bądź konkretnych artykułów opublikowanych na portalach internetowych[54].

DOSTĘPNOŚĆ ZASOBÓW

Następny wykres przedstawia faktyczny dostęp do żądanych zasobów. Do jego stworzenia wykorzystano zarówno dane o liczbie linków działających, otrzymane automatycznie przy użyciu aplikacji, jak i dane o rzeczywistej dostępności zasobów zgromadzone podczas manualnego wyszukiwania informacji pierwotnie niedostępnych.



Wykr. 3. Dostępność zasobów cytowanych w „Rocznikach Bibliotecznych" za lata 1999–2016

Źródło: opracowanie własne.

[54] Ma to miejsce w wypadku „Roczników Bibliotecznych" za 1999 rok, w którym 75% odnośników występuje właśnie w takiej formie. W artykule zatytułowanym *Księgozbiory historyczne w Republice Czeskiej* autorka odwołuje się do strony internetowej dostępnej pod adresem: http://www.lib.cas.cz, która w momencie przeprowadzania badania wciąż była aktywna.

Osiągalność cytowanych dokumentów jest, poza pewnymi wyjątkami, na dość dobrym poziomie — kształtuje się w granicach 70% i więcej. Warto jednak zaznaczyć, że manualne wyszukiwanie treści, do których linki już wygasły, wymaga od odbiorcy wysiłku, a nie każdy użytkownik jest gotowy go podjąć. Sytuacja, kiedy to sam czytelnik wyszukuje sobie źródła konkretnych informacji, nie powinna mieć miejsca w wypadku prawidłowo opracowanych publikacji, ponieważ to w gestii twórcy powinno pozostać takie udokumentowanie cytowanych zasobów, aby można było do nich dotrzeć w bezproblemowy sposób. Taka sytuacja jest jednak możliwa w wypadku zasobów gromadzonych w sieci, ponieważ tylko Internet oraz funkcjonujące w jego ramach narzędzia dają odbiorcom możliwość samodzielnego wyszukiwania treści, na które powołuje się autor danego artykułu. Nie mogłaby ona mieć miejsca przy cytowaniu tradycyjnych publikacji, ponieważ taka operacja jest niewykonalna. Nie byłaby wszakże konieczna, ponieważ dzieła drukowane nie tracą na swojej dostępności tak szybko.

TYPY CYTOWANYCH ŹRÓDEŁ

Podczas opisu źródeł, które cytowali autorzy publikujący w „Rocznikach Bibliotecznych", zwrócono również uwagę na rodzaje przytaczanych źródeł ze względu na ich cechy piśmiennicze i formalne. Wśród nich znalazły się takie kategorie, jak: artykuł, baza danych, bibliografia, czasopismo, dane statystyczne, grafika, hasło encyklopedyczne, katalog, kolekcja, komunikat prasowy, kwestionariusz, materiały konferencyjne, polonika, prezentacja, raport z badań, referat, ustawa, witryna, wynik wyszukiwania, wypowiedź, wystawa i wywiad.

Liczbę odwołań do poszczególnych typów publikacji (cytowanych więcej niż raz) przedstawiono w tabeli 2. W jej ramach zawarto też sumę linków nieaktywnych oraz treści, które są nadal dostępne w innym miejscu sieci.

Tab. 2. Typy cytowanych źródeł

	Liczba cytowań	Liczba nieaktywnych linków	Liczba treści dostępnych po wyszukaniu manualnym
Artykuły	197	68	43
Witryny	99	37	18
Katalogi	10	7	2
Bibliografia	7	1	0
Wypowiedź	6	0	-
Czasopismo	5	0	-
Kolekcje	5	2	2
Wystawa	4	4	0
Prezentacje	3	1	0

	Liczba cytowań	Liczba nieaktywnych linków	Liczba treści dostępnych po wyszukaniu manualnym
Dane statystyczne	3	0	-
Hasła encyklopedyczne	3	1	0
Bazy danych	3	1	1
Materiały konferencyjne	2	1	1
Polonika	2	2	2
Raporty z badań	2	0	-

Źródło: opracowanie własne.

Najczęściej cytowanymi przez naukowców źródłami informacji były artykuły oraz witryny. Pierwsze z wymienionych to dokumenty elektroniczne mające tytuł i autora. Spośród pierwotnie niedziałających 176 linków aż 68 stanowiły odesłania do tego rodzaju utworów. Jednakże dzięki wspomnianym elementom opisu bibliograficznego udało się odszukać 43 teksty, na które powoływali się badacze. Kolejną najliczniejszą kategorią cytowanych dzieł okazały się witryny. Odnośniki je identyfikujące miały przekierować do portali WWW, a dokładniej do ich poszczególnych podstron. Niemniej jednak w tym wypadku dużo trudniej było dotrzeć do odpowiednich zamienników, czyli treści prezentujących informacje, do których odsyłali autorzy „Roczników Bibliotecznych". Wynikało to głównie z faktu, że treści publikowane w sieci, niemające konkretnego autora czy tytułu, są bardzo trudne do zidentyfikowania. Dodatkowo żądane materiały mają płynny charakter i po na przykład przebudowie danej witryny nie można być pewnym, czy w określonej zakładce znajdują się te same informacje, do których wcześniej odsyłał dany link. W takiej sytuacji dużo szybciej może dojść do bezpowrotnej utraty danych.

Inne typy źródeł były cytowane raczej sporadycznie. W kilkunastu wypadkach nie udało się ich zidentyfikować. Problem ten wynikał z niewystarczających informacji zawartych w treści artykułu lub samym opisie bibliograficznym konkretnego źródła.

Stopień dostępności (czyli możliwość odczytania treści ukrytych pod odnośnikami) do poszczególnych rodzajów publikacji jest porównywalny. Ze zgromadzonych danych wynika, iż nie można tylko na podstawie kategorii przywoływanego dzieła przewidzieć długości życia hiperłącza — jej poziom w wypadku zarówno najczęściej cytowanych artykułów, jak i witryn czy haseł encyklopedycznych jest podobny i wynosi około 65%.

Podsumowując, należy wskazać, że początkowo nieaktywnych pozostawało 37% dokumentów (w wyliczeniach wzięto pod uwagę liczbę martwych linków zmniejszoną o hiperłącza uznane pierwotnie za nieaktywne przez program MS

Excel, mimo że można było uzyskać do nich wgląd). Manualne wyszukiwanie żądanych zasobów pozwoliło zmniejszyć tę liczbę do 17%. Należy w tym miejscu również zaznaczyć, że prawidłowo opracowany opis bibliograficzny dokumentów, które zostały zdigitalizowane, umożliwia nie tylko ponowne odszukanie tych publikacji, lecz także korzystanie z materiałów oryginalnych w ich pierwotnej postaci. Z punktu widzenia archiwizacji, co widać na podstawie wyników przeprowadzonego badania, bardzo istotne pozostaje zatem dbanie o trwałość treści o charakterze *born digital*.

Biorąc pod uwagę liczbę badanych odsyłaczy (372) oraz ich rozpiętość w czasie, osiągnięte wyniki wydają się bardzo satysfakcjonujące. Należy jednak podkreślić, że czas niezbędny do manualnego odszukania poszczególnych dokumentów był bardzo długi, a sam proces dość skomplikowany. Jak już zaznaczono, nie każdy czytelnik/użytkownik informacji jest gotów na takie poświęcenie.

PODSUMOWANIE I WNIOSKI

Przeprowadzona analiza potwierdziła zasygnalizowane na wstępie problemy związane z trwałością dostępności zasobów elektronicznych. Internet i funkcjonujące w jego ramach struktury z upływem czasu podlegają znacznym przekształceniom. To jeden z powodów, dla których odsyłacze będące lokalizatorami konkretnych treści sieciowych mogą przekształcić się w nieaktywne linki (nazywane też „martwymi"). Przedstawione badanie nie wykazało bezpośredniej zależności między datą opublikowania dokumentu elektronicznego a trwałością jego odnośnika. Warto jednak zauważyć, że im odnośnik „młodszy", tym większe prawdopodobieństwo odszukania odpowiednich zamienników (czyli wciąż dostępnych treści prezentujących informacje, na które powoływali się autorzy artykułów z „Roczników Bibliotecznych"). Jest ono także dużo większe w wypadku prawidłowo utworzonych przypisów zawierających odnośniki do cytowanych publikacji (zawierających niezbędne elementy opisu bibliograficznego). Ważnym czynnikiem jest również bezbłędne zapisanie/odtworzenie adresu URL w dokumentach cytujących źródła elektroniczne. To właśnie dzięki tym elementom można było zmniejszyć stopień niedostępności do informacji przywoływanych przez autorów tekstów publikowanych w „Rocznikach Bibliotecznych".

Przeprowadzone badania potwierdzają, że najważniejszą cechą zasobów internetowych pozostaje ich dostępność. Nawet jeśli czytelnik nie jest w stanie w jednej chwili przyswoić wszystkich dostarczanych mu treści, to ich odpowiednie archiwizowanie umożliwia powrót do zasobów w innym czasie. Nie powinniśmy zatem zbyt łatwo akceptować tego, iż publikacje, nawet te tworzone przez nas (o prywatnym charakterze), po prostu znikają. Dlatego musimy je przechowywać i udostępniać z tą samą pieczołowitością, która przez lata towarzyszyła kolek-

cjonowaniu oraz upowszechnianiu materiałów piśmienniczych w ich tradycyjnej formie, jaką jest książka w ramach bibliotek. Należy więc budować, popularyzować i wspierać inicjatywy pozwalające na ochronę informacji internetowej, jak również zapewniające jej odpowiednią trwałość.

BIBLIOGRAFIA

About the Internet Archive, Internet Archive, http://www.archive.org/about/about.php [dostęp: 24.11.2018].

Berkeley E.C., *Giant Brains or Machines That Think*, New York 1961, https://monoskop.org/images/b/bc/Berkeley_Edmund_Callis_Giant_Brains_or_Machines_That_Think.pdf [dostęp: 9.11.2018].

Berners-Lee T., Masinter L., McCahill M., *Uniform Resource Locators (URL)*, grudzień 1994, https://tools.ietf.org/html/rfc1738 [dostęp: 16.02.2018].

Bhaskar M., *The Content Machine: Towards a Theory of Publishing from the Printing Press to the Digital Network*, London 2016.

Definicje repozytorium, Baza Wiedzy Politechniki Warszawskiej, http://repo.bg.pw.edu.pl/index.php/pl/informacje-o-repozytorium-o-rep/definicje-repozytorium [dostęp: 9.11.2018].

Derfert-Wolf L., *Archiwizacja Internetu — wprowadzenie i przegląd wybranych inicjatyw*, „Biuletyn EBIB" 2012, nr 1 (128), http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf [dostęp: 24.11.2018].

Dublin Core, Biblioteka Narodowa, https://www.bn.org.pl/dla-bibliotekarzy/normy,-formaty,-standardy/metadane/dublin-core [dostęp: 9.11.2018].

Fajfer A., Imiołek-Stachura K., Januszko-Szakiel A., Patela R., Piwko-Łętek A., Sadlik O., Stachura L., *Trwała ochrona zasobów cyfrowych — podstawowe pojęcia*, „Biuletyn EBIB" 2014, nr 9 (154), http://open.ebib.pl/ojs/index.php/ebib/article/view/311/481 [dostęp: 9.11.2018].

Free ebooks — Project Gutenberg, Projekt Gutenberg, http://www.gutenberg.org/ [dostęp: 9.11.2018].

Gallica, https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop [dostęp: 9.11.2018].

Gmerek K., *Archiwa internetowe po obu stronach Atlantyku — Internet Archive, Wayback Machine oraz UK Web Archive*, „Biuletyn EBIB" 2012, nr 1 (128), http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf [dostęp: 24.11.2018].

Gmiterek G., *Archiwum Internetowe — czy możliwa jest archiwizacja zasobów sieci?*, Biblioteki.org, 25.08.2010, http://www.biblioteki.org/artykuly/Archiwum_Internetowe_czy_mozliwa_jest_archiwizacja_zasobow_sieci-.html [dostęp: 24.11.2018].

Gmiterek G., *Długoterminowa archiwizacja zasobów cyfrowych*, http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.ojs-doi-10_17951_rh_2013_35_213/c/1157-954.pdf [dostęp: 24.11.2018].

Góralska M., *Imperatyw pamięci i akceptacja zapomnienia w epoce cyfrowej*, [w:] *Nauka o informacji w okresie zmian: informatologia i humanistyka cyfrowa*, red. B. Sosińska-Kalata, M. Przastek-Samokowa, Z. Wiorogórska, Warszawa 2016.

Historia Stron, http://historiastron.pl/ [dostęp: 26.11.2018].

Hofmokl J., Tarkowski A., Bednarek-Michalska B., Siewicz K., Szprot J., *Przewodnik po otwartej nauce*, https://repin.pjwstk.edu.pl/files/Przewodnik_po_otwartej_nauce.pdf [dostęp: 23.12.2019].

Hrvatski arhiv weba, http://haw.nsk.hr/ [dostęp: 24.11.2018].

Internet Archive, https://archive.org/ [dostęp: 24.11.2018].

Jankowska M.A., *Biblioteki akademickie — trendy dotyczące zasobów elektronicznych*, 2008, http://www.library.put.poznan.pl/konf_idn/art/4_3.pdf [dostęp: 24.11.2018].

Januszko-Szakiel A., *Archiwistyka cyfrowa. Długoterminowa ochrona dziedzictwa nauki i kultury*, Warszawa 2017.
Karczewska A., *„Wszystko dla ludu przez lud". Popularyzacja czytelnictwa wśród chłopów w Królestwie Polskim w deklaracjach, programie i działalności obozu narodowego w latach 1886–1905*, „Roczniki Biblioteczne" 52, 2008, s. 31–45.
Krzak-Weiss K., *Jeszcze jeden polski sygnet o emblematyczych koneksjach (czyli kilka uwag o znaku firmowym Szymona Kempiniego)*, „Roczniki Biblioteczne" 52, 2008, s. 3–13.
List of Web archiving initiatives, Wikipedia, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 24.11.2018].
Nahotko M., *Metadane. Sposób na uporządkowanie Internetu*, Kraków 2004.
O Polonie, Polona, https://polona.pl/page/o-polonie/ [dostęp: 9.11.2018].
PANDORA, https://pandora.nla.gov.au/ [dostęp: 24.11.2018].
Paradowski D., *Digitalizacja piśmiennictwa*, Warszawa 2010, https://www.bn.org.pl/download/document/1342175805.pdf [dostęp: 6.01.2019].
Parkoła T., *Długoterminowe przechowywanie cyfrowego dziedzictwa kulturowego*, „Biuletyn EBIB" 2014, nr 9 (154), http://open.ebib.pl/ojs/index.php/ebib/article/view/303/477 [dostęp: 9.11.2018].
Portuguese Web Archive, https://arquivo.pt/?l=en [dostęp: 24.11.2018].
Roczniki Biblioteczne, http://rbibl.wuwr.pl/ [dostęp: 13.03.2019].
Spis treści „Roczników Bibliotecznych" z roku 2008, Roczniki Biblioteczne, http://rocznikibiblioteczne.ibi.uni.wroc.pl/nr-52-2008/ [dostęp: 9.03.2019].
Suber P., *A Very Brief Introduction to Open Access*, http://legacy.earlham.edu/~peters/fos/brief.htm [dostęp: 9.11.2018].
UK Web Archive, British Library, https://www.bl.uk/collection-guides/uk-web-archive [dostęp: 25.11.2018].
UKWA, https://www.webarchive.org.uk/en/ukwa/index [dostęp: 24.11.2018].
UKWA About us, UKWA, https://www.webarchive.org.uk/en/ukwa/info/about [dostęp: 24.11.2018].
Usługi powszechnej archiwizacji, PLATON, http://www.platon.pionier.net.pl/online/archiwizacja.php [dostęp: 9.03.2019].
Webarchiv, https://www.webarchiv.cz/ [dostęp: 3.01.2019].
Wilkowski M., *Od osobistej archiwistyki cyfrowej do edukacji medialnej*, „Biuletyn EBIB" 2014, nr 6 (151), http://open.ebib.pl/ojs/index.php/ebib/article/view/274/436 [dostęp: 9.11.2018].
Witczak D., Sobkowiak K., *Problemy przechowywania danych cyfrowych w bibliotekach*, „Elektroniczne Czasopismo Biblioteki Głównej Uniwersytetu Pedagogicznego w Krakowie" 2014, nr 5, http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-e270bf28-b322-47fb--b4a9-8a59456498c1/c/BiE_nr_5_2014_A5_w2.pdf [dostęp: 6.01.2019].
Zdziebłowski S., *Ekspert: potrzeba archiwizowania zasobów internetowych to realny problem*, Nauka w Polsce, 11.01.2018, http://naukawpolsce.pap.pl/aktualnosci/news%2C27860%2Cekspert--potrzeba-archiwizowania-zasobow-internetowych-realny-problem.html [dostęp:26.11.2018].

MARTA TOMALSKA

DURABILITY OF ONLINE INFORMATION BASED ON THE ACCESSIBILITY OF THE WEBSITES CITED IN *ROCZNIKI BIBLIOTECZNE* FROM 1999 TO 2016

Summary

The article deals with the issue of effective striving for constant availability of information published on the Internet in the context of the systems used for this purpose and initiatives that save electronic documents from oblivion. The author attempted to illustrate the considerations regarding the necessity of archiving resources with use of a scientific experiment. As part of it, source material in the form of links placed in footnotes of the content published in the journal *Roczniki Biblioteczne* (Library yearbooks) in the years 1999–2016 was analyzed. The experiment allowed the author to check the frequency of quoted digital publications (372 times), the functioning of the links referring to them (63%), as well as the availability of online resources (83%). It shows how important it is to properly archive web content.

KEY WORDS: durability of information, archiving electronic documents, links, accessibility of digital resources, *Roczniki Biblioteczne*