Bartłomiej Konopa

Naczelna Dyrekcja Archiwów Państwowych, UMK w Toruniu

Wartościowanie i selekcja zasobów WWW – przegląd praktyk

Wartościowanie i selekcja to oczywisty i nieodłączny element praktyki wielu rodzajów archiwów. Pozwalają one wybrać te materiały, które posiadają największą wartość kulturową lub historyczną, a także odpowiedzieć na takie utrudnienia jak ograniczona przestrzeń magazynowa. Zadania te są realizowane również przez funkcjonujące na całym świecie od połowy lat 90. XX w. archiwa webowe. Obecnie istnieją liczne inicjatywy gromadzące zawartość WWW, które realizują swoją misję w różnym zakresie i skali. Część z nich zainteresowana jest zabezpieczaniem cyfrowej części narodowego dziedzictwa kulturowego, inne zachowaniem materiałów wytworzonych przez administracją rządową albo indywidualnych użytkowników Internetu. Mnogość projektów oraz obranych przez nie celów doprowadziło do powstania różnorodnych praktyk związanych z selekcją i wartościowaniem tego rodzaju źródeł.

Podstawowe pytanie jakie należy postawić w niniejszym tekście powinno brzmieć: jak w przypadku archiwów Webu wyglądają wartościowanie i selekcja interesujących je źródeł? Aby móc na nie odpowiedzieć należy najpierw przybliżyć główne obszary, z których wynika konieczność zaimplementowania tych czynności. Należy także zwrócić uwagę na podstawowe problemy, które mogą zmaterializować się w trakcie wartościowania i selekcji zasobów WWW. Umożliwi to przejście do omówienia praktyk stosowanych przez wybrane archiwa webowe, ponieważ są one bezpośrednio powiązane z powyższymi czynnikami. Problematykę selekcji i wartościowania zasobów WWW należy zaprezentować na trzech etapach funkcjonowania inicjatywy je gromadzących. W celu wskazania tych etapów oraz przedstawienia ich we właściwym kontekście wykorzystany zostanie Model Cyklu Życia Archiwizacji Webu (ang. Web Archiving Lifecycle Model)[1]. Za źródło informacji na temat obowiązujących praktyk będą służyły dostępne online przepisy oraz wewnętrzne regulacje, a także opracowania i informacje

[1] M. Bragg, K. Hanna, *Web Archiving Lifecycle Model*, Archive-It marzec 2013, https://archive-it.org/blog/learn-more/publications/web-archiving-life-cycle-model [dostęp 18.07.2022].

udostępniane przez te archiwa webowe. Na koniec możliwe będzie zapytanie o potencjalne zastosowanie omawianych rozwiązań w przypadku powołania polskiego archiwum WWW.

Kwestie wartościowania i selekcji były już podejmowane w literaturze poświęconej archiwom webowym. Wskazać można tu m.in. na artykuł Juliena Masanès *Web Archiving Methods and Approaches: A Comparative Study*, w którym przybliża on i porównuje działalność wybranych inicjatyw, w tym wprowadzane przez nie rozwiązania w zakresie selekcji zasobów WWW[2]. Problematykę wartościowania poruszył m.in. Ed Summers w tekście *Appraisal Talk in Web Archives* opublikowanym w 2020 r., gdzie przedstawione zostały rezultaty wywiadów przeprowadzonych z pracownikami inicjatyw archiwizujących Web odpowiedzialnymi za wartościowanie i selekcję[3]. Warto przywołać także artykuł *Bots, Seeds and People: Web Archives as Infrastructure* przygotowany przez tego autora we współpracy z Ricardo Punzalanem, w którym poruszono również te zagadnienie, jednak w kontekście stosowanych przez archiwa webowe rozwiązań technologicznych[4]. Zagadnienia dotyczące selekcji i wartościowania podejmował także autor niniejszego tekstu, m.in. przy omawianiu działalności narodowych archiwów WWW[5], strategii selektywnej[6] czy wpływu procesu archiwizacji na powstający w jej trakcie zasób[7].

Omawianie problematyki wartościowania zasobów WWW przez archiwa webowe należy rozpocząć od uzasadnienia jej konieczności. Problem ten nie jest obcy archiwistom, ponieważ wartościowanie i selekcja stanowią nieodłączny element ich codziennej praktyki. Dążąc do zachowania zasobów pochodzących z World Wide Web, podobnie jak w przypadku tradycyjnej dokumentacji w formie analogowej lub cyfrowej, należy przyjąć, że zachowanie wszystkich dostępnych w nim treści jest aktualnie niemożliwe. Implikuje to konieczność wyboru tego, która część tych materiałów powinna zostać zarchiwizowana, a która nie. Pozwala się to zgodzić z jednym z pionierów europejskiej archiwistyki webowej Julienem Masanès, który odnośnie do praktyk archiwów WWW stwierdził, że każda archiwizacja jest w jakimś

[2] J. Masanès, *Web Archiving Methods and Approaches: A Comparative Study*, Library Trends, t. 54, 2005, nr 1, DOI: 10.1353/lib.2006.0005, s. 72-90.
[3] E. Summers, *Appraisal Talk in Web Archives*, Archivaria, 2020, nr 89, s. 70-103, https://archivaria.ca/index.php/archivaria/article/view/13733 [dostęp 18.07.2022].
[4] E. Summers, R. Punzalan, *Bots, Seeds and People: Web Archives as Infrastructure*, [w:] *CSCW '17: proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Nowy York 2017, DOI: 10.1145/2998181.2998345, s. 821–834.
[5] B. Konopa, *Archiwizacja Webu w Europie – narodowe archiwa Sieci*, Archeion, t. 121, 2020, DOI: 10.4467/26581264ARC.20.016.12973, s. 445-465.
[6] Idem, *Strategia selektywna jako narzędzie w archiwizacji Webu. Analiza wybranych przykładów*, Archiwa – Kancelarie – Zbiory, t. 11(13), 2020, DOI: 10.12775/AKZ.2020.004, s. 97-118.
[7] Idem, *Reborn digital i black box – wpływ procesu archiwizacji na zasób archiwów Webu*, Archiwa – Kancelarie – Zbiory, t. 10(12), 2019, DOI: 10.12775/AKZ.2019.008, s. 147-168.

stopniu selektywna[8]. Należy zaznaczyć, że ostateczny efekt archiwizacji jest wypadkową wielu czynników, a decyzje wynikające z wartościowania i selekcji to tylko ich część. Istotny wpływ ma również architektura oraz funkcjonowanie Webu, stosowane metody, narzędzia oraz zasoby instytucji podejmującej się archiwizacji, w tym jej kadry i możliwości techniczne oraz prawne.

W przypadku Webu konieczność wartościowania i selekcji jego zasobów wynika nie tylko i nie zawsze z ograniczonej przestrzeni dyskowej oraz mocy obliczeniowej serwerów, które są wykorzystywane na potrzeby archiwizacji zasobów WWW. Niemniej liczba stron internetowych i innych obiektów publikowanych w World Wide Web oraz ich rozmiar stanowią istotną przeszkodę dla inicjatyw, które archiwizują je na szeroką skalę. Warto zaznaczyć także, że określenie tych wartości jest praktycznie niemożliwe oraz nie istnieje kompletny katalog lub indeks stron WWW[9]. Istotny problem stanowi też wysoka dynamika zmian, którą cechuje się Web. Cały czas pojawiają się nowe strony, natomiast poprzednie są zmieniane lub usuwane. Wyliczany na przestrzeni lat 1997-2003 czas życia pojedynczej strony internetowej wahał się od 44 do 100 dni[10], natomiast obecnie dla witryny internetowej, a więc zbioru powiązanych ze sobą stron WWW, może on wynosić około 2 lat i 7 miesięcy[11]. W kontekście zmian należy zwrócić uwagę na ich widoczność. Część z nich może być łatwo dostrzeżona od strony klienta, np. przez użytkownika WWW, m.in. w postaci przekierowania na nowy adres lub błędów protokołu HTTP z serii 4xx i 5xx. Istnieje jednak zjawisko opisywane jako *content drift*, czyli takie przeobrażenia danej strony, które mogą być trudniejsze do wychwycenia[12]. Przykładem tego mogą być artykuły na serwisach informacyjnych, które są publikowane najpierw jako lakoniczna notka, a następnie po krótkim czasie są rozbudowywane.

Istotną kwestią, którą należy uwzględnić w trakcie wartościowania, są prawa autorskie oraz ochrona danych osobowych. Witryny internetowe składają się z różnorodnych elementów, od projektu ich układu, poprzez bogaty wachlarz multimediów, po publikowane teksty, które mogą być przedmiotami prawa autorskiego. Organizacja decydując się na archiwizację jakichś zasobów webowych musi posiadać mandat prawny lub wprowadzić rozwiązania, które taką

[8] J. Masanès, *Selection for Web Archives*, [w:] *Web Archiving*, red. J. Masanès, , Berlin – Nowy York 2006, s. 76.
[9] *How Big Is The Internet? Hint: Probably A Lot Bigger Than You Think*, Starry 29.07.2019, https://starry.com/blog/inside-the-internet/how-big-is-the-internet [dostęp 18.07.2022].
[10] N. Taylor, *The Average Lifespan of a Webpage*, The Signal 08.11.2011, https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/ [dostęp 18.07.2022].
[11] A. Crestodina, *What Is the Average Website Lifespan? 10 Factors In Website Life Expectancy*, Orbit Media Studios 25.04.2017, https://www.orbitmedia.com/blog/website-lifespan-and-you/ [dostęp 18.07.2022].
[12] M. Klein et al., *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot*, PLoS ONE, t. 9, 2014, nr 12, DOI: 10.1371/journal.pone.0115253, s. 2-3.

działalność umożliwią. Sytuacja wygląda podobnie w przypadku danych osobowych. Archiwizacja Webu może być potraktowana jako ich przetwarzanie, w związku z czym wymaga ona odpowiedniego umocowania w prawie. Warto tu przywołać także podejście właścicieli popularnych portali społecznościowych do zbieranych przez nie danych, które znacząco utrudnia archiwizację poprzez ograniczanie dostępu do zawartości tych serwisów[13].

Jako ostatnie warto przywołać architekturę Webu oraz jego ciągły rozwój, które mogą generować zarówno przeszkody, jak i możliwości w zakresie archiwizacji. Podejmując się tego zadania należy uwzględnić cyfrowość WWW, w tym komunikację serwer-klient, budowę składni dokumentu HTML oraz jego multimedialność i hipertekstową strukturę[14]. Pod uwagę należy wziąć także zmiany technologiczne jakie on przechodzi oraz nowe rozwiązanie i elementy, które są do niego wprowadzane. Do klasycznych przykładów zasobów, których gromadzenie jest w znacznym stopniu utrudnione lub niemożliwe, zaliczają się m.in. wycofana w 2020 r. technologia Flash, JavaScript, streamingi czy media społecznościowe[15]. W trakcie wartościowania oraz selekcji zasobów pochodzących z WWW trzeba brać pod uwagę, m.in. ograniczenia narzędzi do ich wyszukiwania oraz późniejszej archiwizacji.

Omówione powyżej obszary pokazują, że konieczność wartościowania i selekcji ma zróżnicowane przyczyny i mogą one mieć istotny wpływ na różne elementy tych procesów. Warto tu zwrócić uwagę na wybrane problemy, które mogą pojawić się w trakcie ich realizacji. Część z nich została już wspomniana w poprzednich akapitach, jednak warto je podkreślić, ponieważ praktyczne rozwiązania, które zostaną omówione w dalszej części, stanowią po części próbę odpowiedzi na napotykane ograniczenia.

W pierwszej kolejności można przywołać utrudnienia wynikające z tego, jak funkcjonuje i rozwija się Web. Będzie to między innymi przywoływana już ilość zachodzących w nim zmian i towarzyszący im jego ciągły wzrost, ponieważ wymaga to rozbudowywania infrastruktury technicznej. Zaliczyć do nich można także nieustające publikowanie w nim nowych treści lub zmienianie lub usuwanie dawnych, które zachodzą w niezwykle dużym tempie, przez co nie jest możliwe zarejestrowanie ich wszystkich. Istotną rolę odgrywa także ewolucja stosowanych technologii, których produkty mogą stanowić przeszkodę dla

[13] Catherine C. Marshall, Frank M. Shipman, *On the Institutional Archiving of Social Media*, [w:] *JCDL '12: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, New York 2012, DOI: 10.1145/2232817.2232819, s. 1-10.
[14] N. Brügger, *The Archived Web. Doing History in the Digital Age*, Cambridge 2018, s. 23–30.
[15] D. Major, D. Gomes, *Web Archives Preserve Our Digital Collective Memory*, [w:] *The Past Web. Exploring Web Archives*, red. D. Gomes, E. Demidova, J. Winters, T. Risse, Cham 2021, DOI: 10.1007/978-3-030-63291-5, s. 17.

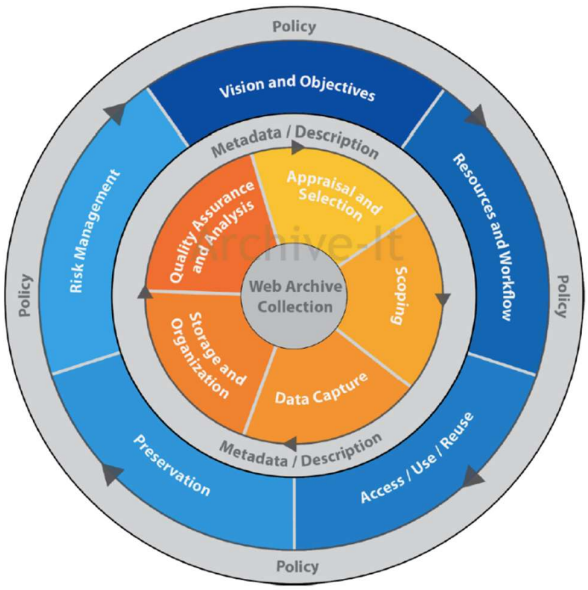
oprogramowania wykorzystywanego w trakcie archiwizacji lub ją uniemożliwiać. W trakcie projektowania narzędzi do gromadzenia i przechowywania zasobów Webu nie jest możliwe przewidzenie przyszłych zmian, a jedynie reagowanie na te, które już zaszły. Problemem może być także zwiększający się rozmiar plików publikowanych w WWW, który ponownie wymagać będzie chociażby zwiększonej przestrzeni dyskowej do przechowywania. Utrudnienie może stanowić również brak jasnych granic w Webie, np. w trakcie wyznaczania jego narodowego czy tematycznego wycinka. Mogą pojawić się także pytania o to, kiedy i w jakiej części archiwizować dane zasoby. Na koniec warto przywołać także brak lub ograniczenia dostępu do pewnych treści, które mogą wynikać m.in. z ich usunięcia, awarii czy wymogu logowania lub wykonania innej nietrywialnej czynności.

Kolejna grupa ograniczeń może wynikać z obowiązujących przepisów prawa. Tak jak zostało to zaznaczone wcześniej witryna internetowa oraz jej poszczególne elementy mogą być przedmiotem prawa autorskiego, w związku z czym ich archiwizacja może wymagać odpowiednich podstaw prawnych lub uprawnień. Dodatkowe utrudnienie może stanowić ustalenie autora lub właściciela danych treści oraz nawiązanie z nim kontaktu, np. w celu uzyskania zgody na wykonanie kopii należącej do niego witryny. Ponadto problem może wynikać z tego, że część zasobów jest publikowana w WWW z naruszeniem praw autorskich lub innych przepisów. Należy uwzględnić także to, że zasoby webowe mogą zawierać dane osobowe, które także podlegają ochronie prawnej. Ponadto nie każdy twórca lub inny użytkownik WWW musi godzić się na archiwizację wytworzonych przez niego treści i należy uwzględnić jego prawo do zapomnienia.

Ostatnim obszarem, który może generować utrudnienia w trakcie wartościowania i selekcji zasobów WWW jest zawartość publikowanych w nim treści. W związku z czym należy przede wszystkim wypracować kryteria oraz inne narzędzia, które pozwolą zachować zasoby odpowiadające zakresowi działalności archiwum. Problem może stanowić nagromadzenie treści o nikłej wartości informacyjnej, takie jak np. spam czy scam. Z drugiej strony warto zadać pytanie, czy zjawiska uznane współcześnie za nieistotne lub szkodliwe nie wzbudzą zainteresowania badaczy. Wybór zagadnień, a następnie poszczególnych witryn i innych zasobów do archiwizacji, tak aby odpowiedzieć na potrzeby obecnych i przyszłych użytkowników, jest bardzo wymagającym zadaniem.

Z powyższego zestawienia problemów oraz wcześniej omówionych obszarów jasno wynika, że wartościowanie i selekcja zasobów WWW muszą odbywać się na różnych etapach oraz uwzględniać szereg różnorodnych czynników. Aby pokazać, jak wygląda w tym zakresie

praktyka archiwów webowych korzystne będzie posłużenie się Modelem Cyklu Życia Archiwizacji WWW. Został on wypracowany przez pracowników Archive-It, płatnej usługi archiwizacyjnej oferowanej przez Internet Archive, we współpracy z wybranymi jej użytkownikami. Pomimo tego, że były to instytucje działające na relatywnie niedużą skalę, zaproponowany model można zastosować także do pozostałych inicjatyw zajmujących się gromadzeniem i przechowywaniem zasobów pochodzących z Webu. Model został przedstawiony w formie graficznej (zob. Ilustracja 1) w postaci koła podzielonego na pięć zasadniczych obszarów. Pierwszy z nich to otaczająca pozostałe Polityka, a więc zbiór strategicznych decyzji danej organizacji dotyczących archiwizacji. Na następny obszar składają się 5 kluczowych zagadnień związanych z powoływaniem oraz zarządzaniem archiwum. Są to: Wizja i Cele, Zasoby i Przepływ Pracy, Dostęp i Wykorzystanie, Przechowywanie oraz Zarządzanie Ryzykiem. Kolejny poziom modelu to Metadane oraz Opis, które otaczając niższą warstwę, składająca się z 5 czynności wykonywanych w codziennej pracy archiwów: Wartościowanie i Selekcja, Ustalanie Zakresu, Gromadzenie Danych, Przechowywanie i Organizowanie oraz Sprawdzenie Jakości i Analiza. W centrum wizualizacji modelu znajduje się kolekcja zasobów zgromadzonych w trakcie archiwizacji, będąca efektem powyższych działań i decyzji z nimi związanych [16]. Na potrzeby omówienia zagadnień dotyczących wartościowania i selekcji zasobów WWW wykorzystane zostaną następujące elementy opisanego powyżej modelu: Polityka, Wizja i Cele oraz Wartościowanie i Selekcja.



Ilustracja 1 Model Cyklu Życia Archiwizacji Webu, źródło: M. Bragg, K. Hanna, *Web Archiving Lifecycle Model*, Archive-It marzec 2013, https://archive-it.org/blog/learn-more/publications/web-archiving-life-cycle-

[16] M. Bragg, K. Hanna, *Web Archiving Lifecycle Model*.

model/ [dostęp 18.07.2022]

Jako pierwszy obszar należy omówić poziom Polityki instytucji, ponieważ od pierwszej strategicznej przez nią decyzji o zachowaniu jakichś zasobów WWW rozpoczyna się uznanie ich znaczenia jako materiałów wartych zabezpieczenia i długotrwałego przechowywania. Wiąże się to z przypisaniem im wartości historycznej lub kulturowej, niezależnie od tego czy zostaną potraktowane jako materiał archiwalny, biblioteczny egzemplarz obowiązkowy lub inna forma dziedzictwa kulturowego. Przyznanie zawartości WWW istotnej roli we współczesnym świecie dostrzec można w dokumentach publikowanych przez międzynarodowe organizacje, np. w ogłoszonej przez UNESCO w 2003 roku Karcie Ochrony Cyfrowego Dziedzictwa (ang. Charter on the Preservation of the Digital Heritage)[17] czy w Zaleceniach Komisji Europejskiego z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych[18]. Widoczne jest ono również w normatywach poszczególnych państw, które umożliwiają odpowiednim instytucjom archiwizację Webu, np. poprzez objęcie ich przepisami o egzemplarzu obowiązkowym. Przykłady takich przepisów można znaleźć m.in. w Danii[19] lub Zjednoczonym Królestwie[20]. Pokazuje to także praktyka działających na całym świecie archiwów Webu, które gromadzą i przechowują różnorodne wycinki WWW. Pierwsze taki inicjatywy zostały uruchomione w połowie lat 90, XX w., a więc mowa tu o ponad dwudziestopięcioletnich doświadczeniach. Motywacje do podjęcia decyzji o archiwizacji World Wide Web mogą być różne i będą wynikać m.in. z szerszych celów lub zadań danej organizacji oraz będą przekładać na dalsze działania, w tym określenie celów jakich będzie dążyć powołane archiwum. Niemniej na poziomie polityki istotna jest sama decyzja o konieczności zachowania określonego wycinka zasobów Webu, ponieważ wiąże się ona z uznaniem ich istotnej wartości.

Kolejnym ze obszarów wybranych z Modelu Cyklu Życia Archiwizacji Webu są Wizja i Cele, a więc decyzje instytucji archiwizującej dotyczące tego co ma być gromadzone i przechowywane, jaki ma przynieść to efekt i jakimi metodami ma zostać on osiągnięty. Wiąże się to z określeniem celów tej instytucji oraz tego, jak wpisują się one w powierzone jej

[17] UNESCO, *Charter on the Preservation of the Digital Heritage*, 17.10.2003, UNESDOC Digital Library, https://unesdoc.unesco.org/ark:/48223/pf0000179529 [dostęp 18.07.2022].
[18] Zalecenie Komisji z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych, Dz.U. L 283 z 29.10.2011, s. 39-45, http://data.europa.eu/eli/reco/2011/711/oj [Dostęp 18.07.2022].
[19] LOV nr 1439 af 22/12/2004 Lov om pligtaflevering af offentliggjort materiale, Retsinformation, https://www.retsinformation.dk/eli/lta/2004/1439 [dostęp 18.07.2022].
[20] UK Statutory Instruments 2013 Nr. 777 The Legal Deposit Libraries (Non-Print Works) Regulations 2013, https://www.legislation.gov.uk/uksi/2013/777/contents/made [dostęp 18.07.2022].

zadania[21]. Dotychczasowa działalność archiwów Webu pozwala na dokonanie ich podziału według tego, jaką część zasobów WWW decydują się archiwizować:

- całość publicznie dostępnego WWW – takie podejście do archiwizacji jest wyjątkiem, ponieważ wskazać można na dwie główne inicjatywy, które je realizują – Internet Archive[22] oraz Common Crawl[23]. Pierwsza z nich to pionier archiwizacji Webu i największe istniejące do tej pory archiwum webowe, które dąży do zachowania jak największej części zasobów WWW i zapewnienia do nich uniwersalnego dostępu. Common Crawl prowadzi zbliżoną działalność, jednak zdecydowanie krócej oraz na mniejszą skalę i jest nastawione na udostępnianie danych na potrzeby badań i rozwoju;
- narodowy wycinek Webu – jest to popularne rozwiązanie stosowane w wielu państwach, które sprowadza się do archiwizowania tej części zasobów WWW, które powiązane są z danym narodem. Relacja ta może być ustalana na podstawie różnych czynników, takich jak np. domena krajowa najwyższego poziomu czy język. Wśród obecnie funkcjonujących narodowych archiwów webowych można zauważyć dwa sposoby gromadzenia takich zasobów, których zastosowanie wynika z możliwości jakimi dysponuje dana instytucja. Może to być kompleksowe zabezpieczanie narodowego WWW poprzez regularną archiwizację całej domeny krajowej oraz innych domen, jeżeli takowe istnieją, oraz stosowanie innych metod takich jak gromadzenie tematyczne lub dokumentowanie wydarzeń. Przykłady archiwów realizujących to podejście można znaleźć m.in. w Zjednoczonym Królestwie (UK Web Archive[24]), Danii (Netarkivet[25]), Portugalii (Arquivo.pt[26]), Chorwacji (Hrvatski Arhiv Weba[27]),

[21] M. Bragg, K. Hanna, *Web Archiving Lifecycle Model*.
[22] A. Ben-David, A. Amram, *The Internet Archive and the socio-technical construction of historical facts*, Internet Histories, t. 2, 2018, nr 1-2, DOI: 10.1080/24701475.2018.1455412, s. 179-181.
[23] *Frequently Asked Questions*, Common Crawl, https://commoncrawl.org/big-picture/frequently-asked-questions/ [dostęp 18.07.2022].
[24] *Collection guides. UK Web Archive*, British Library, https://www.bl.uk/collection-guides/uk-web-archive [dostęp 18.07.2022].
[25] *Netarkivet*, Det Kgl. Bibliotek, https://www.kb.dk/en/find-materials/collections/netarkivet [dostęp 18.07.2022].
[26] *Crawling web content*, Arquivo.pt, https://sobre.arquivo.pt/en/help/crawling-and-archiving-web-content/ [dostęp 18.07.2022].
[27] *About HAW*, HAW Croatian Web Archive, https://haw.nsk.hr/en/about-haw/ [dostęp 18.07.2022].

Australii (Australian Web Archiv[28]) czy Singapurze (Web Archive Singapore[29]). Drugie ze używanych rozwiązań w archiwizacji narodowego fragmentu WWW ogranicza się do wykorzystania wyłącznie metod selektywnej archiwizacji. Taki sposób działania stosowany był w pierwszych latach funkcjonowania przywoływanych archiwów w Zjednoczonym Królewskie, Chorwacji oraz Australii (wówczas jako PANDORA [30]), a obecnie jest on używany w Niderlandach (Webarchief van Nederland[31]), Nowej Zelandii (New Zealand Web Archive[32]) czy Japonii (Web Archiving Project[33]);

- tematyczny wycinek WWW – jest to grupa rozwiązań stosowanych przez liczne archiwa webowe, które przeważnie działają na mniejszą skalę od omawianych powyżej inicjatyw o charakterze narodowym. Wśród różnorodnych praktyk można wyróżnić kilka grup. Jednym z częstych podejść do archiwizacji tematycznej jest gromadzenia zasobów wytwarzanych przez instytucje rządowe i samorządowe. Liczną grupę takich inicjatyw znaleźć można w Stanach Zjednoczonych Ameryki (np. End of Term Web Archive[34], CyberCemetery[35], North Carolina State Government Web Site Archives and Access Program[36] oraz archiwum webowe Biblioteki Kongresu USA[37]), a także w Australii (The Australian Government Web Archive, obecnie włączone do Australian Web Archive[38]) czy w Zjednoczonym Królestwie (UK Government Web Archive[39]). Kolejnym rodzajem archiwów WWW możliwym do wydzielenia są inicjatywy

[28] *Australian Web Archive*, National Library of Australia, https://www.nla.gov.au/collections/building-our-collections/australian-web-archive [dostęp 18.07.2022].
[29] *Frequently Asked Questions*, Web Archive Singapore, https://eresources.nlb.gov.sg/webarchives/faq [dostęp 18.07.2022].
[30] *PANDORA Overview*, PANDORA Australia's Web Archive, http://pandora.nla.gov.au/overview.html [dostęp 18.07.2022].
[31] *Web archiving*, KB Nationale Bibliotheek, https://www.kb.nl/en/about-us/expertise/web-archiving [dostęp 18.07.2022].
[32] *New Zealand Web Archive*, National Library, https://natlib.govt.nz/collections/a-z/new-zealand-web-archive [dostęp 18.07.2022].
[33] *Archiving Internet Information*, National Diet Library, Japan, https://www.ndl.go.jp/en/collect/internet/index.html [dostęp 18.07.2022].
[34] *Project Background*, The End of Term Web Archive, http://eotarchive.cdlib.org/background.html [dostęp 18.07.2022].
[35] *CyberCemetery*, UNT Digital Library, https://digital.library.unt.edu/explore/collections/GDCC/ [dostęp 18.07.2022].
[36] *About the Web Site Archives and Access Program*, NC State Government Web Site Archives and Access Program, https://webarchives.ncdcr.gov/about.html [dostęp 18.07.2022].
[37] *About This Program*, Library of Congress, https://www.loc.gov/programs/web-archiving/about-this-program/ [dostęp 18.07.2022].
[38] *Australian Web Archive*, National Library of Australia.
[39] *About the UK Government Web Archive*, The National Archives, https://www.nationalarchives.gov.uk/webarchive/about-the-uk-government-web-archive/ [dostęp 18.07.2022].

gromadzące zasoby na potrzeby badań naukowych, do których zaliczyć można Latin American Web Archiving Project prowadzony na Uniwersytecie Teksasu w Austin[40], obecnie nieaktywne Digital Archive for Chinese Studies działające w ramach Uniwersytetu w Heidelbergu[41]. Podobna działalność jest prowadzona przez Bibliotekę Uniwersytetu Columbia [42] czy Bibliotekę Bodleiańską Uniwersytetu[43]. Innym stosowanym podejściem jest archiwizowanie materiałów związanych z istotnymi tematami lub wydarzeniami. Realizuje je m.in. Content Development Working Group działająca w ramach International Internet Preservation Consortium [44] czy wymieniana już Biblioteka Kongresu USA. Wskazać można także na indywidualne rozwiązania, np. na działalność artystycznej grupy non-profit Rhizome, która gromadzi i zabezpiecza sztukę Internetu[45] czy ArchiveTeam, nieformalny kolektyw, który skupia się przede wszystkim na archiwizacji treści wytwarzanych przez użytkowników, a pochodzącej z serwisów i portali, które mają zostać usunięte z WWW[46].

Przywołane powyżej przypadki pokazują znaczne zróżnicowanie tego, co poszczególne archiwa WWW oraz inne instytucje obierają za cel własnej działalności. W tym miejscu należy zapytać z czego wynika taki, a nie inny zakres archiwizacji prowadzonej przez daną instytucję? Przede wszystkim jest to efekt rozpoznania potrzeby zabezpieczenia danego wycinka Webu oraz uznania go za dziedzictwo kulturowe wymagające długotrwałego przechowywania. Widoczne jest to na przykładzie Internet Archive oraz ArchiveTeam, ale stanowi to także motywację dla pozostałych istniejących inicjatyw. Wpływ mają na to również statutowe zadania, które powierzono danym instytucjom, np. archiwom lub bibliotekom. W wielu krajach są one zobligowane do zabezpieczania dziedzictwa związanego z państwem, w którym funkcjonują, a swoją działalność rozszerzają na jego nowe formy w postaci zasobów WWW, aby jak najpełniej wypełniać swoja misję. Nie bez znaczenia są również obowiązujące przepisy. Tak jak zostało to podkreślone wcześniej archiwizacja WWW wymaga także odpowiednich

[40] *Latin American Web Archiving Project*, Latin American Network Information Center, http://lanic.utexas.edu/project/archives/ [dostęp 18.07.2022].
[41] *About DACHS*, Bereichsbibliothek Ostasien, https://www.zo.uni-heidelberg.de/boa/digital_resources/dachs/about_en.html [dostęp 18.07.2022].
[42] *About the program*, Columbia University Libraries, https://library.columbia.edu/collections/web-archives/about_program.html [dostęp 18.07.2022].
[43] *Web Archives: Home*, Bodleian Libraries, https://libguides.bodleian.ox.ac.uk/web-archives [dostęp 18.07.2022].
[44] *Content Development Working Group*, International Internet Preservation Consortium, https://netpreserve.org/about-us/working-groups/content-development-working-group/ [dostęp 18.07.2022].
[45] *About*, Rhizome/ArtBase, https://artbase.rhizome.org/wiki/About [dostęp 18.07.2022].
[46] *Philosophy*, Archive Team, https://wiki.archiveteam.org/index.php/Philosophy [dostęp 18.07.2022].

podstaw prawnych lub wprowadzenia odpowiednich mechanizmów, które pozwolą gromadzić dane materiały. Wpływ legislacji widoczny jest na przykładzie archiwów, które posiadają możliwość archiwizacji Webu w oparciu o przepisy dotyczące egzemplarza obowiązkowego, co otwiera drogę do regularnego gromadzenia całej domeny krajowej. W przypadku braku takich zapisów działalność archiwum webowego jest znacząco ograniczona i często wymaga pozyskania wcześniejszej zgody właściciela danej witryny na jej archiwizację. Wpływ na zakres działalności posiadają również zasoby jakimi dysponuje dana inicjatywa. Archiwizacja Webu jest kosztownym przedsięwzięciem, wymagającym odpowiedniego zaplecza technicznego, specjalistycznego oprogramowania oraz kompetentnych kadr, a zatem ich deficyt może ograniczać taką działalność.

Obrany przez dane archiwum zakres archiwizacji bezpośrednio determinuje jego przyszły zasób i posiada wpływ na to, co zostanie zarchiwizowane, a co nie. Zobrazować można to dość oczywistym przykładem: inicjatywa gromadząca zasoby pochodzące z australijskiej krajowej domeny nie będzie zainteresowana domeną polską lub portugalską, o ile ich zasoby nie dotyczą Australii lub Australijczyków. Ponadto obranie szerokiego zakresu archiwizacji, to jest całości dostępnego publicznie WWW lub jego narodowego wycinka, skutkować będzie gromadzeniem niezwykle obszernych zasobów. Warto zwrócić uwagę na wprowadzane rozwiązania, które służą realizacji celów, które stawiają przed sobą archiwa WWW. Przede wszystkim jest to odpowiednie zaimplementowanie jednej z dwóch strategii archiwizacji. W przypadku dużych archiwów jest to najczęściej połączenie gromadzenia masowego i selektywnego, co umożliwia kompleksowe zachowanie narodowego fragmentu WWW, natomiast mniejsze inicjatywy ograniczają się do stosowanie archiwizacji selektywnej. Na potrzeby realizacji obu tych strategii opracowywane są szczegółowe kryteria, które stosowane są w trakcie selekcji zasobów, niezależnie od tego czy odbywa się ona manualnie czy automatycznie[47].

Następnym etapem wybranym z Modelu Cyklu Życia Archiwizacji Webu jest Wartościowanie i Selekcja, a więc zbiór czynności związanych z realizacją omawianych powyżej celów i zamierzeń[48]. Wiąże się on z opracowaniem oraz wykorzystywaniem przywołanych powyżej kryteriów. Zgodnie z zaprezentowanymi wcześniej czynnikami, które archiwum WWW musi uwzględnić w kontekście selekcji zasobów, stosowane kryteria można podzielić na trzy grupy: merytoryczne, formalne oraz techniczne. Pierwsza z nich odnosi się do

[47] B. Konopa, *Strategia selektywna jako narzędzie w archiwizacji Webu*, s. 101-103.
[48] M. Bragg, K. Hanna, *Web Archiving Lifecycle Model*.

zawartości witryny internetowej, publikowanych na niej treści, ich znaczenia i wartości informacyjnej, relewantności czy popularności. Kolejna opiera się na takich elementach jak domena, na której zarejestrowana jest witryna, jej autor lub właściciel, jego lokalizacja, a także język czy kwestie prawne. Ostatnia grupa to kryteria techniczne, na którą składają się m.in. format i rozmiar plików, ich ilość lub stosowane rozwiązania technologiczne.

Kryteria merytoryczna to kluczowa grupa kryteriów stosowanych w trakcie selekcji zasobów webowych, ponieważ pozwalają one stwierdzić czy dana witryna internetowa lub inny materiał wpisuje się w zakres działalności danego archiwum. Przykłady takich rozwiązań dostrzec można m.in. w wytycznych stosowanych w Bibliotece Kongresu USA. Wskazują one, że archiwizowane zasoby powinny wpisywać się w szersze obszary działalności Biblioteki, a także odpowiadać potrzebom Kongresu oraz badaczy, zawierać unikatowe i jakościowe informacje, treści o charakterze naukowym czy być powiązane z innymi zbiorami tej instytucji[49]. Zbliżone podejście deklaruje Biblioteka Narodowa Australii, która w ramach selektywnego gromadzenia australijskiego Webu jest zainteresowana zasobami, które dotyczą istotnych wydarzeń społecznych, kulturalnych czy politycznych oraz zjawisk, będących przedmiotem debaty publicznej, a także treściami naukowymi czy dotyczącymi mniejszości etnicznych[50]. Kwestię unikatowości oraz wartość informacyjnej, a także tematykę związaną z Chorwacją uwzględnia Chorwackie Archiwum Webu[51]. Podobne kryteria stosowane są podczas tworzenia kolekcji zasobów WWW budowanych przez członków International Internet Preservation Consortium w ramach Content Development Working Group. Włączane do nich materiały powinny być m.in. użyteczne do przyszłych badań naukowych oraz posiadać ponadnarodowe znaczenie[52]. Popularność wśród użytkowników WWW uwzględnia chociażby kolektyw ArchiveTeam[53] oraz duński Netarkivet[54].

Istotną rolę odgrywają także kryteria formalne, na podstawie których ustala m.in. przynależność wycinka Webu do danego narodu. Najczęściej w tym celu wykorzystuje się domenę krajową najwyższego poziomu oraz domeny regionalne, co jest praktyką stosowaną w

[49] *Library Of Congress Collections Policy Statements Supplementary Guidelines. Web Archiving*, czerwiec 2022, s. 4-5, https://www.loc.gov/acq/devpol/webarchive.pdf [dostęp 18.07.2022].
[50] *What we collect*, National Library of Australia, https://www.nla.gov.au/about-us/corporate-documents/policy-and-planning/collection-development-policy/what-we-collect [dostęp 18.07.2022].
[51] K. Holub, I. Rudomino, *A decade of web archiving in the National and University Library in Zagreb*, materiały z konferencji IFLA WLIC 2015, Kapsztad (RPA), 11-20 sierpnia 2015, s. 3-4, http://library.ifla.org/1092/1/090-holub-en.pdf [dostęp 18.07.2022].
[52] *Content Development Working Group*, International Internet Preservation Consortium.
[53] *Philosophy*, Archive Team.
[54] S. Schostag, E. Fønss-Jørgensen, *Webarchiving: Legal Deposit of Internet in Denmark*. A Curatorial Perspective, Microform & Digitization Review, t. 41, 2012, nr 3-4, DOI: 10.1515/mir-2012-0018, s. 112-115.

tych archiwach narodowych webowych, które mają prawne podstawy do ich archiwizacji. Przykładem tego może być UK Web Archive, które poza regularnym gromadzeniem domeny .uk, swoją działalnością obejmuje też domeny .scot, .wales, .cymru and .london[55]. Biblioteka Narodowa Francji w ramach archiwizacji WWW poza domeną francuską .fr, zabezpiecza zasoby z domen należących do terytoriów zamorskich Francji[56]. Innym stosowanym kryterium jest lokalizacja serwera hostującego daną witrynę. Wykorzystywane jest ono m.in. w Netarkivet[57] oraz UK Web Archive. Do grupy kryteriów formalnych można zaliczyć także język czy narodowość autora, które stosuje Chorwackie Archiwum Webu[58]. Kwestia twórcy zasobów odgrywa także istotną rolę także w archiwach gromadzących materiały wytwarzane przez władze państwowe oraz samorządowe. Za przykład posłużyć program Congressional & Federal Government Web Harvest, prowadzony przez amerykańską agencję National Archives and Records Administration, w ramach którego gromadzone są witryny należące do Senatu i Izby Reprezentantów, ich władz, poszczególnych członków oraz działających przy nich komisji[59]. Projekt brytyjskich The National Archives – UK Government Web Archive wynika z ich statutowej działalności, a więc zabezpieczaniem państwowego zasobu archiwalnego, w związku z czym dąży do pełnego gromadzenia witryn wytworzonych przez centralne władze Zjednoczonego Królestwa[60].

Kryteria techniczne pełnią odmienną rolę od pozostałych omówionych powyżej i wynikają z barier i problemów technicznych, które napotykają archiwa WWW. Spowodowane jest to niedoskonałością oprogramowania wykorzystywanego przez te inicjatywy oraz ich ograniczonymi możliwościami. Zaliczyć można do nich też występujące w Webie ograniczenia dostępu dla robotów archiwizujących, np. konieczność zalogowania się, aby uzyskać dostęp do części zasobów, na co zwraca uwagę m.in. UK Government Web Archive[61]. Problemy tego typu mogą wynikać także z czasowego lub stałego braku dostępu do danej witryny, np. w wyniku awarii serwera lub odpowiedniej konfiguracji Robots Exclusion Protocol, a więc ustawień odpowiedzialnych za wpuszczanie na witrynę robotów indeksujących. Warto

[55] *Collection guides. UK Web Archive*, British Library.
[56] S. Aubry, *Web Archives as a New Library Service: the Experience of the National Library of France*, LIBER Quarterly, t. 20, 2010, nr 2, DOI: 10.18352/lq.7987 s. 182.
[57] S. Schostag, E. Fønss-Jørgensen, *Webarchiving*, s. 111.
[58] K. Holub, I. Rudomino, *A decade of web archiving*, s. 3-4.
[59] *Congressional & Federal Government Web Harvests*, The National Archives, https://www.webharvest.gov/ [dostęp 18.07.2022].
[60] *Operational Selection Policy OSP27. UK Central Government Web Estate*, The National Archives czerwiec 2014, s. 5-6, https://cdn.nationalarchives.gov.uk/documents/information-management/osp27.pdf [dostęp 18.07.2022].
[61] Ibidem, s. 8.

zauważyć, że podejście do tego ostatniego jest dwojakie i część archiwów je respektuje, np. Chorwackie Archiwum Webu[62], lub ignoruje, np. duński Netarkivet[63]. Na podstawie podobnych kryteriów z archiwizacji części zasobów rezygnuje Biblioteka Kongresu. Ponadto zaznacza, że poza jej zasięgiem znajdują się m.in. publikowane w WWW bazy danych, streamingi oraz media społecznościowe, a także zasoby Głębokiego Webu, a więc te, do których dotarcie za pomocą linków lub wyszukiwarek jest utrudnione[64]. Podobne podejście można dostrzec także na przykładach UK Web Archive[65] czy Netarkivet[66]. Możliwości techniczne archiwum mogą także ograniczać zakres gromadzonych zasobów, np. Chorwackie Archiwum Webu, które ze względu na nie wystarczającą przestrzeń dyskową rezygnowało z pobierania plików powyżej 100MB[67].

Na koniec należy zadać pytanie o to, które z omawianych celów, rozwiązań lub praktyk stosowanych przez archiwa webowe możliwe byłoby do wprowadzenia w Polsce? Należy tu odnotować fakt, że aktualnie w Polsce działa jedynie jeden projekt archiwizacji zasobów WWW, a mianowicie Archiwum Społeczne Polskiego Webu prowadzone przez Marcina Wilkowskiego[68]. Brak systemowych rozwiązań takich jak np. kompleksowa archiwizacja domeny krajowej, które funkcjonują w innych rejonach świata, wynika m.in. z funkcjonujących ograniczeń prawnych oraz braku przygotowania. Wracając jednak do postawionego pytania można wskazać na kilka działań, które byłyby warte rozważenie w kontekście gromadzenia zasobów pochodzących z WWW w Polsce. Pierwszym koniecznym krokiem jest uznanie ich wartości kulturowej i historycznej oraz wynikających z nich korzyści dla badań naukowych lub zabezpieczenia i udokumentowania zjawisk ważnych społecznie. Może to dotyczyć całości polskiego Webu lub jakiegoś jego wycinka. Za inspirację mogą posłużyć tu projekty archiwizujące zasoby wytwarzane przez instytucje publiczne, takie jak UK Government Web Archive czy japoński Web Archiving Program, lub takie, które ze względu na brak odpowiednich podstaw prawnych selektywnie gromadzą fragmenty WWW związane z danym narodem, np. te działające w Niderlandach czy Nowej Zelandii. Posiadają one wypracowane kryteria merytoryczne i formalne umożliwiające realizację postawionych przed nim celów, a

[62] K. Holub, I. Rudomino, *A decade of web archiving*, s. 7.
[63] *About collecting internet material*, Det Kgl. Bibliotek, https://www.kb.dk/en/find-materials/collections/netarkivet/about-collecting-internet-material [dostęp 18.07.2022].
[64] *Library Of Congress Collections Policy Statements*, s. 5.
[65] *Frequently asked questions*, UK Web Archive, https://www.webarchive.org.uk/en/ukwa/info/faq [dostęp 18.07.2022].
[66] S. Schostag, E. Fønss-Jørgensen, *Webarchiving*, s. 119-120.
[67] K. Holub, I. Rudomino, *A decade of web archiving*, s. 7-8.
[68] *Informacje o projekcie*, Archiwum Społeczne Polskiego Webu, https://aspw.pl/informacje [dostęp 18.07.2022].

które mogłyby zostać zaimplementowane w ich polskim odpowiedniku. Pamiętać należy również o uwzględnieniu możliwości technicznych oraz o problemach, które mogą wynikać z tego jak funkcjonuje współczesny Web, ponieważ mogą one wymusić rezygnację z archiwizacji części zasobów.

Podsumowując należy przypomnieć, że archiwa webowe, podobnie jak innego rodzaju archiwa, w swojej działalności muszą uwzględniać wartościowanie i selekcję zasobów, które zamierzają gromadzić. Konieczność ta wynika m.in. z wciąż powiększającego się rozmiaru WWW oraz ilości zachodzących w nim zmian, przez co zachowanie całości publikowanych w nim treści jest niemożliwe. Ponadto wymuszają ją rozwój technologiczny Webu oraz wprowadzanie do niego nowych elementów, na które narzędzia stosowane w archiwizacji mogą nie być przygotowane. Uwzględnienia wymagają również obowiązujące przepisy prawne dotyczące m.in. ochrony danych osobowych czy własności intelektualnej. Istotne jest także określenie wartości oraz relewantności danych zasobów.

Wartościowanie i selekcje w przypadku archiwów WWW odbywa się na trzech etapach ich funkcjonowania. Za pierwszy uznać można decyzję o uznaniu wartości kulturowej zasobów i treści publikowanych w World Wide Web lub jakiejś ich części. Następnym krokiem jest sprecyzowanie celu i zakresu działania takiej inicjatywy. Do najczęściej występujących modeli zaliczyć można archiwizację narodowego wycinka Webu lub zasobów powiązanych tematycznie lub formalnie, np. wytwarzanych przez władze państwowe. Nie należy jednak zapominać o inicjatywach, które swoją działalność próbują prowadzić jak najszerzej, np. o Internet Archive.

Obrany przez archiwum cel będzie wpływał na metody wykorzystywane w jego realizacji. Ich istotny komponent stanowią kryteria selekcji stosowane przez poszczególne inicjatywy gromadzące zasoby Webu. Dotychczas stosowane kryteria podzielić na trzy grupy: merytoryczne, formalne oraz techniczne. Kryteria merytoryczne odnoszą się do treści zasobów WWW, przez co pełnią ważną rolę w selektywnej strategii archiwizacji, ponieważ umożliwiają ocenę wartości danych zasobów. Do drugiej grupy zaliczają się takie czynniki jak domena, na której witryna jest zarejestrowana, lokalizacja jej hostowania czy język, które są przydatne przy wyznaczaniu narodowego fragmentu WWW. Kryteria techniczne są natomiast odpowiedzią na ograniczenia na jakie napotkać może infrastruktura lub narzędzia archiwum oraz są dostosowywane do ich możliwości. Na zakończenie warto zaznaczyć, że rozwiązania stosowane i rozwijane od prawie 30 lat przez archiwa webowe działające na całym świecie mogą zostać zaimplementowane w przyszłym polskim archiwum WWW.

Bibliografia

About collecting internet material, Det Kgl. Bibliotek, https://www.kb.dk/en/find-materials/collections/netarkivet/about-collecting-internet-material [dostęp 18.07.2022].

About DACHS, Bereichsbibliothek Ostasien, https://www.zo.uni-heidelberg.de/boa/digital_resources/dachs/about_en.html [dostęp 18.07.2022].

About HAW, HAW Croatian Web Archive, https://haw.nsk.hr/en/about-haw/ [dostęp 18.07.2022].

About the program, Columbia University Libraries, https://library.columbia.edu/collections/web-archives/about_program.html [dostęp 18.07.2022].

About the UK Government Web Archive, The National Archives, https://www.nationalarchives.gov.uk/webarchive/about-the-uk-government-web-archive/ [dostęp 18.07.2022].

About the Web Site Archives and Access Program, NC State Government Web Site Archives and Access Program, https://webarchives.ncdcr.gov/about.html [dostęp 18.07.2022].

About This Program, Library of Congress, https://www.loc.gov/programs/web-archiving/about-this-program/ [dostęp 18.07.2022].

About, Rhizome/ArtBase, https://artbase.rhizome.org/wiki/About [dostęp 18.07.2022].

Archiving Internet Information, National Diet Library, Japan, https://www.ndl.go.jp/en/collect/internet/index.html [dostęp 18.07.2022].

Aubry S., *Web Archives as a New Library Service: the Experience of the National Library of France*, LIBER Quarterly, t. 20, 2010, nr 2, DOI: 10.18352/lq.7987 s. 179–-199.

Australian Web Archive, National Library of Australia, https://www.nla.gov.au/collections/building-our-collections/australian-web-archive [dostęp 18.07.2022].

Ben-David A., Amram A., *The Internet Archive and the socio-technical construction of*

Bragg M., Hanna K., *Web Archiving Lifecycle Model*, Archive-It marzec 2013, https://archive-it.org/blog/learn-more/publications/web-archiving-life-cycle-model [dostęp 18.07.2022].

Brügger N., *The Archived Web. Doing History in the Digital Age*, Cambridge 2018.

Catherine C. Marshall, Frank M. Shipman, *On the Institutional Archiving of Social Media*, [w:] *JCDL '12: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, New York 2012, DOI: 10.1145/2232817.2232819, s. 1-10.

Collection guides. UK Web Archive, British Library, https://www.bl.uk/collection-guides/uk-web-archive [dostęp 18.07.2022].

Congressional & Federal Government Web Harvests, The National Archives, https://www.webharvest.gov/ [dostęp 18.07.2022].

Content Development Working Group, International Internet Preservation Consortium, https://netpreserve.org/about-us/working-groups/content-development-working-group/ [dostęp 18.07.2022].

Crawling web content, Arquivo.pt, https://sobre.arquivo.pt/en/help/crawling-and-archiving-web-content/ [dostęp 18.07.2022].

Crestodina A., *What Is the Average Website Lifespan? 10 Factors In Website Life Expectancy*, Orbit Media Studios 25.04.2017, https://www.orbitmedia.com/blog/website-lifespan-and-you/ [dostęp 18.07.2022].

CyberCemetery, UNT Digital Library, https://digital.library.unt.edu/explore/collections/GDCC/ [dostęp 18.07.2022].

Frequently Asked Questions, Common Crawl, https://commoncrawl.org/big-picture/frequently-asked-questions/ [dostęp 18.07.2022].

Frequently asked questions, UK Web Archive, https://www.webarchive.org.uk/en/ukwa/info/faq [dostęp 18.07.2022].

Frequently Asked Questions, Web Archive Singapore, https://eresources.nlb.gov.sg/webarchives/faq [dostęp 18.07.2022].

historical facts, Internet Histories, t. 2, 2018, nr 1-2, DOI: 10.1080/24701475.2018.1455412, s. 179-201.

Holub K., Rudomino I., *A decade of web archiving in the National and University Library in Zagreb*, materiały z konferencji IFLA WLIC 2015, Kapsztad (RPA), 11-20 sierpnia 2015, http://library.ifla.org/1092/1/090-holub-en.pdf [dostęp 18.07.2022].

How Big Is The Internet? Hint: Probably A Lot Bigger Than You Think, Starry 29.07.2019, https://starry.com/blog/inside-the-internet/how-big-is-the-internet [dostęp 18.07.2022].

Informacje o projekcie, Archiwum Społeczne Polskiego Webu, https://aspw.pl/informacje [dostęp 18.07.2022].

Klein M. et al., *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot*, PLoS ONE, t. 9, 2014, nr 12, DOI: 10.1371/journal.pone.0115253.

Konopa B., *Archiwizacja Webu w Europie – narodowe archiwa Sieci*, Archeion, t. 121, 2020, DOI: 10.4467/26581264ARC.20.016.12973, s. 445-465.

Konopa B., *Reborn digital i black box – wpływ procesu archiwizacji na zasób archiwów Webu*, Archiwa – Kancelarie – Zbiory, t. 10(12), 2019, DOI: 10.12775/AKZ.2019.008, s. 147-168.

Konopa B., *Strategia selektywna jako narzędzie w archiwizacji Webu. Analiza wybranych przykładów*, Archiwa – Kancelarie – Zbiory, t. 11(13), 2020, DOI: 10.12775/AKZ.2020.004, s. 97-118.

Latin American Web Archiving Project, Latin American Network Information Center, http://lanic.utexas.edu/project/archives/ [dostęp 18.07.2022].

Library Of Congress Collections Policy Statements Supplementary Guidelines. Web Archiving, czerwiec 2022, https://www.loc.gov/acq/devpol/webarchive.pdf [dostęp 18.07.2022].

LOV nr 1439 af 22/12/2004 Lov om pligtaflevering af offentliggjort materiale, Retsinformation, https://www.retsinformation.dk/eli/lta/2004/1439 [dostęp 18.07.2022].

Major D., Gomes D., *Web Archives Preserve Our Digital Collective Memory*, [w:] *The Past Web. Exploring Web Archives*, red. D. Gomes, E. Demidova, J. Winters, T. Risse, Cham 2021, DOI: 10.1007/978-3-030-63291-5, s. 11-19.

Masanès J., *Selection for Web Archives*, [w:] *Web Archiving*, red. J. Masanès, , Berlin – Nowy York 2006, s. 71-91.

Masanès J., *Web Archiving Methods and Approaches: A Comparative Study*, Library Trends, t. 54, 2005, nr 1, DOI: 10.1353/lib.2006.0005, s. 72-90.

Netarkivet, Det Kgl. Bibliotek, https://www.kb.dk/en/find-materials/collections/netarkivet [dostęp 18.07.2022].

New Zealand Web Archive, National Library, https://natlib.govt.nz/collections/a-z/new-zealand-web-archive [dostęp 18.07.2022].

Operational Selection Policy OSP27. UK Central Government Web Estate, The National Archives czerwiec 2014, https://cdn.nationalarchives.gov.uk/documents/information-management/osp27.pdf [dostęp 18.07.2022].

PANDORA Overview, PANDORA Australia's Web Archive, http://pandora.nla.gov.au/overview.html [dostęp 18.07.2022].

Philosophy, Archive Team, https://wiki.archiveteam.org/index.php/Philosophy [dostęp 18.07.2022].

Project Background, The End of Term Web Archive, http://eotarchive.cdlib.org/background.html [dostęp 18.07.2022].

Schostag S., Fønss-Jørgensen E., *Webarchiving: Legal Deposit of Internet in Denmark*. A Curatorial Perspective, Microform & Digitization Review, t. 41, 2012, nr 3-4, DOI: 10.1515/mir-2012-0018, s. 110-120.

Summers E., *Appraisal Talk in Web Archives*, Archivaria, 2020, nr 89, s. 70-103, https://archivaria.ca/index.php/archivaria/article/view/13733 [dostęp 18.07.2022].

Summers E., Punzalan R., *Bots, Seeds and People: Web Archives as Infrastructure*, [w:] *CSCW '17: proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Nowy York 2017, DOI: 10.1145/2998181.2998345, s. 821–834.

Taylor N., *The Average Lifespan of a Webpage*, The Signal 08.11.2011, https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/ [dostęp 18.07.2022].

UK Statutory Instruments 2013 Nr. 777 The Legal Deposit Libraries (Non-Print Works) Regulations 2013, https://www.legislation.gov.uk/uksi/2013/777/contents/made [dostęp 18.07.2022].

UNESCO, *Charter on the Preservation of the Digital Heritage*, 17.10.2003, UNESDOC Digital Library, https://unesdoc.unesco.org/ark:/48223/pf0000179529 [dostęp 18.07.2022].

Web Archives: Home, Bodleian Libraries, https://libguides.bodleian.ox.ac.uk/web-archives [dostęp 18.07.2022].

Web archiving, KB Nationale Bibliotheek, https://www.kb.nl/en/about-us/expertise/web-archiving [dostęp 18.07.2022].

What we collect, National Library of Australia, https://www.nla.gov.au/about-us/corporate-documents/policy-and-planning/collection-development-policy/what-we-collect [dostęp 18.07.2022].

Zalecenie Komisji z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych, Dz.U. L 283 z 29.10.2011, s. 39-45, http://data.europa.eu/eli/reco/2011/711/oj [Dostęp 18.07.2022].