ARCHEION, T. CXXI
WARSZAWA 2020
ISSN 0066-6041
e-ISSN 2658-1264 DOI 10.4467/26581264ARC.20.016.12973

BARTŁOMIEJ KONOPA
ORCID 0000-0001-9843-5552
bkonopa@doktorant.umk.pl
(Uniwersytet Mikołaja Kopernika w Toruniu, Archiwum Państwowe w Bydgoszczy)

ARCHIWIZACJA WEBU W EUROPIE – NARODOWE ARCHIWA SIECI

Słowa kluczowe: archiwizacja Webu, archiwa Webu, archiwa cyfrowe, witryny Internetowe, badania nad Internetem

Streszczenie

Archiwizacja Webu, czyli działania mające na celu gromadzenie i zachowanie zasobów Sieci, prowadzona jest już od prawie 25 lat. Przez ten czas powstało wiele projektów realizujących to zadanie, a także parę organizacji, takich jak np. International Internet Preservation Consortium, które wspierają jego realizowanie. W artykule zaprezentowano rozwój działań w tym zakresie, a następnie omówiono wnioski z analizy funkcjonowania wybranych europejskich archiwów Sieci o charakterze narodowym, przeprowadzonej w oparciu o publicznie dostępne materiały ich dotyczące. Analiza ta miała na celu zbadanie, w jaki sposób obecnie archiwizowany jest Web w tej części świata. Rozpatrzone zostały trzy główne zagadnienia: gromadzenie, opisywanie i udostępnianie zasobów dawnego WWW. Pierwsze z nich obejmuje zakres archiwizacji, a więc określenie tego, jakie materiały jej podlegają, a także wykorzystywanych w tym celu strategii, z których wynika ukształtowanie zbiorów. Drugie dotyczy stosowanych metadanych i innych elementów służących przekazaniu informacji na temat tego, co zostało w jej trakcie zgromadzone. Ostatni element analizy obejmuje zakres udostępniania zasobów archiwalnego WWW, występujące ograniczenia i ich przyczyny, a także wykorzystywane do tego narzędzia. W trakcie badań zainteresowano się również używanym przez poszczególne projekty oprogramowaniem. Uzyskane wyniki pozwalają stwierdzić, że model archiwum Sieci został wypracowany, a działalność analizowanych inicjatyw w Europie jest do siebie bardzo zbliżona.

BARTŁOMIEJ KONOPA
ORCID 0000-0001-9843-5552
bkonopa@doktorant.umk.pl
(Nicolaus Copernicus University in Toruń, State Archive in Bydgoszcz)

WEB ARCHIVING IN EUROPE – NATIONAL WEB ARCHIVES

Key words: Web archiving, Web archives, digital archives, websites, Internet studies

Abstract

Web archiving, that is activities aimed at collecting and preserving Web resources, has been carried out for almost 25 years. During this time, many projects have been created to fulfill that task, as well as several organizations, such as the International Internet Preservation Consortium, that support it implementation. The article presents the development of activities in this area, and then presents the conclusions of the analysis of the functioning of selected European national Web archives, based on publicly available materials concerning them. This analysis was intended to examine how the Web is currently archived in this part of the world. Three main issues were considered: gathering, describing and access to the resources of the former WWW. The first of them covers the scope of archiving, namely determining what materials are subject to it, as well as the gathering strategies used for this purpose, which shape the archival collections. The second concerns the metadata and other elements used to convey information about what was collected during that process. The last element of the analysis includes the scope of access to archival WWW resources, existing restrictions and their causes, as well as the tools used for this. During the research, the author also became interested in the software used in individual projects. The obtained results show that the model of Web archive has been developed and the activities of the analyzed initiatives in Europe are very similar.

Wstęp[1]

Internet odgrywa obecnie szczególną rolę w życiu współczesnych ludzi i jest istotnym medium wymiany informacji. Służy on kulturze, sztuce oraz nauce, a także codziennej pracy i komunikacji. Potencjał Sieci został dość szybko dostrzeżony i w połowie lat 90. XX w. podjęto pierwsze próby jej archiwizacji. Od tego momentu prowadzone są działania mające na celu gromadzenie i zabezpieczenie zasobów Webu. Archiwizacja Sieci, niekiedy mylnie nazywana archiwizacją Internetu[2], polega na wyszukiwaniu materiałów w World Wide Web i wykonywaniu ich zrzutów (eng. „snapshots") za pomocą robota internetowego. Następnie kopie te są przechowywane w specjalnych formatach plików i udostępniane za pomocą odpowiedniego oprogramowania, które ma za zadanie odtworzyć wrażenie korzystania z witryny internetowej w momencie jej archiwizacji. Na całym świecie, w tym w Europie, organizowane są projekty, które mają realizować takie zadania, a gromadzony przez nie zasób już teraz jest wykorzystywany w badaniach naukowych.

Celem analizy, której wyniki zostaną zaprezentowane w niniejszym artykule, było zbadanie w jaki sposób zadanie archiwizacji WWW jest realizowane w Europie. W pierwszej kolejności wymagało to przyjrzenia się rozwojowi tego zjawiska na świecie, a zwłaszcza na Starym Kontynencie. Dotyczyło to zarówno inicjatyw zajmujących się archiwizacją zasobów Webu, jaki i projektów mających na celu rozwój i wspieranie takich działań. Następnie przebadane zostały wybrane archiwa Sieci o charakterze narodowym. Do analizy zebrano informacje dotyczące trzech podstawowych zagadnień: gromadzenia zasobów, a dokładnie zakresu i stosowanych w jego trakcie strategii, ich opisywania oraz udostępniania. Sprawdzano także wykorzystywane oprogramowanie.

Literatura i źródła

Problematyka archiwizacji Webu podejmowana jest w różnego rodzaju publikacjach. Pierwszy podręcznik z tego zakresu został wydany w 2006 r. pod redakcją Juliena Masanès`a[3], a w 2018 ukazała się książka poświęcona zagadnieniu historii Webu, redagowana przez Nielsa Brüggera i Iana Milligana[4]. Warta uwagi jest również publikacja *The Web as History: Using Web Archives to Understand*

[1] Artykuł został przygotowany w ramach grantu badawczego, realizowanego przez Wydział Nauk Historycznych UMK w Toruniu nr 1130-NH Europejskie projekty archiwizacji Internetu – zakres, działalność, stosowane rozwiązania.

[2] Internet to globalna sieć połączeń pomiędzy komputerami, natomiast Web to jedna z jego usług, za pomocą której udostępnia się różne zasoby (np. w postaci witryn WWW), które następnie mogą być archiwizowane, zob. Wikipedia, World Wide Web a Internet, https://pl.wikipedia.org/wiki/World_Wide_Web#World_Wide_Web_a_Internet [dostęp: 7.06.2020].

[3] *Web Archiving*, oprac. i red. J. Masanès, Berlin–Heidelberg 2006.

[4] *The SAGE Handbook of Web History*, oprac. i red. N. Brügger, I. Milligan, Thousand Oaks 2018.

the Past and the Present, w której zebrano liczne badania nad zasobami archiwalnego WWW[5]. Ukazują się również artykuły naukowe[6] oraz raporty[7] podejmujące tę problematykę. Tematyka archiwizacji Sieci była również poruszana na łamach polskiej literatury archiwalnej, wspomnieć tu można chociażby artykuł Filipa Kłębczyka, który ukazał się w „Archiwiście Polskim"[8], oraz Agnieszki Rosy, opublikowany w „Archiwach – Kancelariach – Zbiorach"[9], a także referat Anny Sobczak wydany w 4. tomie pokonferencyjnym „Toruńskich Konfrontacji Archiwalnych"[10]. Archiwizacja Webu pojawiła się także w trakcie piątej edycji tej konferencji[11]. Podsumowanie rozważań nad tą problematyką w Polsce opublikował niedawno Wojciech Woźniak[12].

Dane na temat działalności europejskich archiwów WWW, potrzebne do przeprowadzenia niniejszych rozważań, zostały pozyskane z List of Web archiving initiatives[13] oraz z witryn internetowych bibliotek narodowych, prowadzących takie archiwa. Wykorzystane zostały też przepisy prawne poszczególnych państw regulujące ich działalność, a także dostępne w Sieci materiały promocyjne oraz pokonferencyjne. Część projektów doczekała się artykułów monograficznych, w których zostało szerzej opisane ich funkcjonowanie; dotyczy to

[5] *The Web as History: Using Web Archives to Understand the Past and the Present*, oprac. i red. N. Brügger, R. Schroeder, Londyn 2017, https://www.jstor.org/stable/j.ctt1mtz55k [dostęp: 7.06.2020].

[6] Np.: M.M. Farag, S. Lee, E.A. Fox, *Focused crawler for events*, „International Journal on Digital Libraries" 2018, t. 19, nr. 1, s. 3–19, https://link.springer.com/article/10.1007/s00799-016-0207-1 [dostęp: 7.06.2020].

[7] Np.: M.D. Costea, *Report on the Scholarly Use of Web Archives*, Aarhus 2018, http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf [dostęp: 7.06.2020].

[8] F. Kłębczyk, *Archiwizacja zasobów Internetu – kierunki i wyzwania*, „Archiwista Polski" 2012, nr 3(67), s. 105–112.

[9] A. Rosa, *Human trace on the Internet – the issue of archiving the Web from the point of view of anthropology-oriented archival science*, „Archiwa – Kancelarie – Zbiory" 2015, t. 6(8), s. 193–205, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2018.003 [dostęp: 7.06.2020].

[10] A. Sobczak, *Internet jako globalne archiwum społeczne – rozważania na temat roli Internetu w dokumentowaniu dziejów ludzkości*, [w:] *Nowa archiwistyka – archiwa i archiwistyka w późnonowoczesnym kontekście kulturowym*, Toruńskie Konfrontacje Archiwalne, t. 4, oprac. i red. W. Chorążyczewski, W. Piasek, A. Rosa, Toruń 2014, s. 237–247.

[11] Laboratorium Cyfrowe Humanistyki UW, Toruńskie Konfrontacje Archiwalne i problemy archiwizacji Webu, https://lach.edu.pl/blog/2017/12/11/torunskie-konfrontacje-archiwalne-problemy-archiwizacji-webu/ [dostęp: 07.06.2020].

[12] W. Woźniak, *Archiwizacja Internetu – próba podsumowania dotychczasowych prac i ustaleń*, „Archiwa – Kancelarie – Zbiory" 2015, t. 10(12), s. 75–98, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2019.004 [dostęp: 7.06.2020].

[13] Wikipedia, List of Web archiving initiatives, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 07.06.2020].

m.in. archiwów Webu francuskiego[14], duńskiego[15] oraz chorwackiego[16]. Należy zwrócić uwagę na fakt, iż inicjatywy te upubliczniają mało informacji na temat swojego funkcjonowania, co utrudnia ich dobre poznanie oraz wykorzystanie przez zainteresowanych[17].

Rozwój archiwizacji Webu na świecie

Pierwsze próby gromadzenia danych z Webu w celu jego długotrwałego zachowania podejmowano w Kanadzie w latach 1994–1996. Za początek archiwizacji WWW należy jednak uznać 1996 r., kiedy swoją działalność rozpoczęła fundacja Archiwum Internetu (Internet Archive), która dysponuje obecnie największymi zasobami dawnej Sieci. Podjęcie tych działań pokazuje jak wcześnie zdano sobie sprawę z roli rozwijającego się medium, jakim wówczas był Internet powstały na początku lat 90. XX w. Pomysłodawcą pierwszej takiej inicjatywy był Brewster Kahle, przedsiębiorca internetowy i autor serwisu Alexa zajmującego się badaniem ruchu generowanego pomiędzy witrynami internetowymi. Postanowił on wykorzystać dane pozyskane podczas trwania jego działalności do uruchomienia Archiwum Internetu, które miało stać się bazą do badań nad tym zjawiskiem. Fundacja ta stworzyła podstawowe oprogramowanie używane w archiwizacji Internetu i jest jednym z głównych pomysłodawców International Internet Preservation Consortium (IIPC). W 1996 r., oprócz Internet Archive, powołano dwie podobne inicjatywy: archiwum Webu Australii – PANDORA, oraz Szwecji – Kulturarw3. W następnym roku kolejny projekt uruchomiono w Nowej Zelandii, zaś 3 lata później tego zadania podjęła się Biblioteka Kongresu USA[18].

W pierwszej dekadzie XXI w. pojawiły się dwa specyficzne zjawiska w archiwizacji Sieci: archiwizacja tematyczna oraz narodowe archiwa Sieci. Do pierwszego zaliczyć można kolekcje zasobów webowych powstałe w wyniku działania różnych instytucji i organizacji, które postanowiły zachować jakiś ich

[14] S. Aubry, *Web Archives as a New Library Service: the Experience of the National Library of France*, „LIBER Quarterly" 2010, t. 20, nr 2, s. 179–199, https://www.liberquarterly.eu/articles/10.18352/lq.7987/ [dostęp: 7.06.2020].

[15] S. Schostag, E. Fønss-Jørgensen, *Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective*, „Microform & Digitization Review" 2012, t. 41, nr 3–4, s. 110–120, https://www.degruyter.com/view/journals/mfir/41/3-4/article-p110.xml [dostęp: 7.06.2020].

[16] K. Holub, I. Rudomino, *A decade of web archiving in the National and University Library in Zagreb* (materiały z konferencji IFLA WLIC 2015, Kapsztad, 11–20 sierpnia 2015), s. 1–12, http://library.ifla.org/1092/1/090-holub-en.pdf [dostęp: 7.06.2020].

[17] Problem ten został zauważony w artykule: A. AlSum, M.C. Weigle, M. L. Nelson, H. Van de Sompel, *Profiling web archive coverage for top-level domain and content language*, „International Journal on Digital Libraries" 2014, t. 14, nr 3–4, s. 149, https://link.springer.com/article/10.1007%2Fs00799-014-0118-y [dostęp: 7.06.2020].

[18] R. Schroeder, N. Brügger, *Introduction: The web as history*, [w:] *The Web as History*, s. 6–7.

wycinek związany z konkretnym tematem lub wydarzeniem. Jako przykład przywołać tu można m.in. działalność wspomnianej już Biblioteki Kongresu (posiada ona kolekcje archiwalnych witryn dotyczące Igrzysk Olimpijskich z 2002 r., wojny w Iraku, czy wyboru papieża Benedykta XVI)[19] oraz uruchomienie przez Internet Archive komercyjnej usługi Archive-It, która umożliwia zainteresowanym instytucjom gromadzenie zasobów Webu według ich własnych potrzeb[20]. Wspomnieć należy także o różnych oddolnych inicjatywach, o których w swoim artykule pisał Marcin Wilkowski[21]. W tym czasie zaczęły też powstawać na całym świecie, w tym także w Europie, archiwa Sieci o charakterze narodowym, za które w większości przypadków odpowiedzialne są biblioteki narodowe. Jedne z pierwszych takich projektów powstały m.in. w: Norwegii, Danii, Chorwacji, Czechach czy Wielkiej Brytanii[22].

Kolejne projekty archiwizacji WWW są wciąż inicjowane. W trakcie swoich badań ankietowych – *A survey on web archiving initiatives* – Daniel Gomez, Joâo Miranda oraz Miguel Costa zgromadzili dane na temat 42 takich inicjatyw (33 na podstawie ankiet, 9 na podstawie publicznie dostępnych informacji). Duża grupa pochodziła ze Stanów Zjednoczonych, lecz aż 23 z nich miały miejsce w Europie[23]. Obecnie zaś, według listy prowadzonej na Wikipedii oraz bazującej na wynikach zaprezentowanych we wspomnianej powyżej publikacji, odnotowano 93 inicjatywy zaangażowane w przeprowadzanie i rozwijanie archiwizacji Webu, z czego 45 jest związanych z Europą[24]. Najmłodsze wymienione wśród nich archiwum pochodzi z Belgii i prace nad nim, z udziałem archiwów państwowych Belgii oraz tamtejszej Biblioteki Narodowej, są na ukończeniu[25].

---

[19] Pełen wykaz kolekcji zarchiwizowanych witryn w zasobie Biblioteki Kongresu USA, zob.: Library of Congress, Digital Collections, https://www.loc.gov/collections/?fa=original-format:archived+web+site [dostęp: 7.06.2020].

[20] Archive-It, Archive-It Blog – About us, https://archive-it.org/blog/learn-more/ [dostęp: 7.06.2020].

[21] M. Wilkowski, *Oddolne archiwizacje Internetu jako działania społeczne*, „Archiwa-Kancelarie-Zbiory" 2015, t. 6(8), s. 207–220, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2015.007 [dostęp: 7.06.2020].

[22] R. Rogers, *Periodizing Web Archiving: Biographical, Event-Based, National and Autobiographical Traditions*, [w:] *The SAGE Handbook…*, s. 45–46; R. Schroeder, N. Brügger, op.cit., s. 7–8.

[23] D. Gomez, J. Miranda, M. Costa, A *survey on web archiving initiatives*, [w:] *Research and Advanced Technology for Digital Libraries. International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26–28, 2011. Proceedings*, oprac. i red. S. Gradmann, F. Borri, C. Meghini, H. Schuldt, Berlin 2011, s. 410–413, https://link.springer.com/chapter/10.1007/978-3-642-24469-8_41 [dostęp: 7.06.2020].

[24] Wikipedia, List of Web archiving initiatives, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 7.06.2020].

[25] F. Geereart, S. Soyez, *The first steps towards a Belgian web archive: a federal strategy (materiały z konferencji IIPC Web Archiving Conference 2019, Zagrzeb, 6–7 czerwca 2019)*, http://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-FRIEDEL_GEERAERT__

Realizacja tych projektów napotyka różne problemy i przeszkody, które można podzielić na 3 podstawowe grupy:

- merytoryczne – podejmując się archiwizacji WWW, należy określić jej zakres, wykorzystywane strategie, stosowane kryteria selekcji i metody opisu zasobów,
- prawne – przepisy dotyczące praw autorskich, egzemplarza obowiązkowego lub danych osobowych mają znaczący wpływ na zakres gromadzenia oraz udostępniania,
- techniczne – w celu realizacji postawionych przed projektem celów należy przygotować odpowiednie oprogramowanie oraz sprzęt komputerowy.

Badacze dostrzegli też pewne negatywne zjawisko, polegające na niewielkim wykorzystywaniu zasobów archiwalnego Webu w nauce, wynikające z różnych ograniczeń w dostępie, trudności w ich przeszukiwaniu oraz braku odpowiednich do tego metod. W ostatnich latach sytuacja ta zaczęła jednak ulegać poprawie i studia nad dawną Siecią zaczynają się powoli rozwijać[26].

Archiwa Webu wspierane są przez różnego rodzaju inicjatywy, które zajmują się opracowaniem rozwiązań w tym zakresie. Na tym polu kluczową rolę odgrywa wymienione już wcześniej Internet Archive, które jest autorem najczęściej wykorzystywanego oprogramowania (robot archiwizujący Heritrix oraz technologia Wayback Machine) oraz właścicielem usługi Archive-It. Ponadto wspiera ono narodowe archiwa WWW w gromadzeniu zasobów[27] oraz jest jednym z głównych inicjatorów IIPC, które zostało powołane w 2003 r. W momencie powstania zrzeszało 12 członków, obecnie zaś liczy ich już 57, wśród których znajdują się biblioteki narodowe (w tym polska Biblioteka Narodowa), brytyjskie Archiwum Narodowe, uniwersytety oraz inne organizacje zajmujące się archiwizacją Sieci. Konsorcjum zajmuje się wspieraniem inicjatyw oraz rozwijaniem oprogramowania stworzonego przez Internet Archive, a także organizuje konferencje, warsztaty i grupy robocze, które umożliwiają wymianę wiedzy i doświadczenia[28]. Oprócz IIPC podobną rolę pełniła Internet Memory Foundation (założona w 2004 r. w Amsterdamie jako European Archive Foundation, zakoń-

SEBASTIEN_SOYEZ-The_first_steps_towards_a_Belgian_web_archive-a_federal_strategy.pdf [dostęp: 9.06.2020].

[26] R. Schroeder, N. Brügger, *Introduction: The web...*, s. 9–13; R. Rogers, *Periodizing Web Archiving...*, s. 49–53.

[27] Fundacja Internet Archive wspierała archiwizację m.in. Webu hiszpańskiego (zob.: Biblioteca Nacional de España, History of the collection, http://www.bne.es/en/Colecciones/ArchivoWeb/Historia/index.html [dostęp: 7.06.2020]), francuskiego (zob.: Bibliothèque nationale de France, Archives de l'internet, https://www.bnf.fr/fr/archives-de-linternet [dostęp: 8.06.2020]) oraz irlandzkiego (zob.: National Library of Ireland, Irish Domain Web Archive, https://www.nli.ie/en/irish-domain-web-archive.aspx [dostęp: 8.06.2020]).

[28] International Internet Preservation Consortium, About IIPC, http://netpreserve.org/about-us/ [dostęp: 8.06.2020]; ibidem, IIPC members, http://netpreserve.org/about-us/members/ [dostęp: 8.06.2020].

czyła działalność w 2018 r.), która zajmowała się zarówno archiwizacją Webu, jaki i opracowaniem odpowiednich ku temu narzędzi[29].

Oprócz tego organizowane są różne przedsięwzięcia, które popularyzują wykorzystanie zasobów archiwalnej Sieci w nauce oraz poszukują metod pozwalających na ich efektywne zastosowanie. Jako przykład można wymienić projekt RESAW (A Research Infrastructure for the Study of Archived Web Materials), który funkcjonuje od 2012 r. i zrzesza badaczy m.in. z Danii, Francji, Wielkiej Brytanii oraz Holandii, używających tego rodzaju źródła w swoich badaniach[30]. Zbliżone zakresem działania, lecz mniejsze od nich inicjatywy powołuje się w poszczególnych państwach Europy. Jako przykład podać można projekt BUDDAH (Big UK Domain Data for the Arts and Humanities)[31] czy duński NetLab[32]. Zbliżoną funkcję w Polsce pełnił webArch – pracowania archiwizacji Webu działająca od czerwca 2018 r. przy Laboratorium Cyfrowym Humanistyki (obecnie Centrum Kompetencji Cyfrowych) Uniwersytetu Warszawskiego wspierana przez pracowników i doktorantów Uniwersytetu Mikołaja Kopernika w Toruniu[33].

Narodowe archiwa Webu w Europie

Jak widać z powyższego omówienia, archiwizacja WWW w Europie nie jest zjawiskiem zupełnie nowym i powoli umacnia swoją pozycję wśród innych działań mających na celu zachowanie dziedzictwa kulturowego. Przez 24 lata udało się też wypracować wiele rozwiązań w tym zakresie. Jak już wcześniej wspomniano, na Starym Kontynencie funkcjonuje lub funkcjonowało ponad 40 różnych inicjatyw związanych z tym zagadnieniem. W niniejszym artykule zostaną przedstawione wyniki analizy działalności 14 projektów zajmujących się gromadzeniem, przechowywaniem i udostępnianiem zbiorów dawnej Sieci, które w większości mają charakter ogólnonarodowy (dwa przypadki to mniejszości narodowe w Hiszpanii). Przy wyborze przykładów kierowano się dostępem do dostatecznie aktualnych informacji na ich temat w języku angielskim, dostępnych w witrynach internetowych należących do inicjatyw archiwizacji Webu oraz w artykułach naukowych lub innych materiałach. Podstawowe dane na ich temat zostały przedstawione w tabeli nr 1.

[29] Wikipedia, Internet Memory Foundation, https://en.wikipedia.org/wiki/Internet_Memory_Foundation [dostęp: 8.06.2020].

[30] A Research Infrastructure for the Study of Archived Web Materials, About RESAW, http://resaw.eu/about/ [dostęp: 8.06.2020]; A Research Infrastructure for the Study of Archived Web Materials, Participants http://resaw.eu/participants/ [dostęp: 8.06.2020].

[31] Big UK Domain Data for the Arts and Humanities, Aims and objectives, https://buddah.projects.history.ac.uk/about/aims-and-objectives/ [dostęp: 8.06.2020].

[32] NetLab, Mission, http://www.netlab.dk/netlab/mission/ [dostęp: 8.06.2020].

[33] webArch CKC UW, Pracownia archiwizacji Webu CKC UW, https://webarch.uw.edu.pl/pracownia/ [dostęp: 8.06.2020].

Tabela 1: Wybrane projekty archiwizacji Webu w Europie – opracowanie własne

Nazwa projektu	Adres URL	Kraj pochodzenia	Data założenia	Strategia gromadzenia	Dostęp do zasobów
WebArchiv	https://www.webarchiv.cz/	Czechy	2000	masowa i selektywna	częściowo ograniczony
Archives de l'internet	https://www.bnf.fr/fr/archives-de-linternet	Francja	2002	masowa i selektywna	ograniczony
Hrvatski arhiv weba	http://haw.nsk.hr/en	Chorwacja	2004	masowa i selektywna	otwarty
Vefsafn.is	https://vefsafn.is/	Islandia	2004	masowa i selektywna	otwarty
UK Web Archive	https://www.webarchive.org.uk/	Wielka Brytania	2004	masowa i selektywna	częściowo ograniczony
PADICAT (Patrimoni Digital de Catalunya)	https://www.padicat.cat/	Hiszpania (Katalonia)	2005	masowa i selektywna	częściowo ograniczony
Netarkivet	http://netarkivet.dk/in-english/	Dania	2005	masowa i selektywna	ograniczony
Suomalainen verkkoarkisto	http://verkkoarkisto.kansalliskirjasto.fi/va/	Finlandia	2006	masowa i selektywna	częściowo ograniczony
Ondarenet	http://www.ondarenet.kultura.ejgv.euskadi.eus:8085/ondarenet/	Hiszpania (Kraj Basków)	2007	selektywna	otwarty
Webarchief van Nederland	https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving	Holandia	2007	selektywna	częściowo ograniczony
Arquivo.pt	https://arquivo.pt/?l=en	Portugalia	2008	masowa i selektywna	otwarty
Archivo de la Web Española	http://www.bne.es/en/Colecciones/ArchivoWeb/index.html	Hiszpania	2009	masowa i selektywna	częściowo ograniczony
Eesti veebiarhiivi	http://veebiarhiiv.digar.ee/	Estonia	2010	masowa i selektywna	częściowo ograniczony
Web Archive	https://archive-it.org/home/nli	Irlandia	2011	masowa i selektywna	częściowo ograniczony

Omawiając działalność projektów archiwizacji Webu, w pierwszej kolejności należy zwrócić uwagę na proces gromadzenia, a dokładnie na jego zakres oraz wykorzystane w jego trakcie strategie. Wszystkie z analizowanych inicjatyw budują swoje zasoby w oparciu o kryterium narodowe, a więc swoim zasięgiem obejmują np. sieć francuską, czeską, portugalską itd. O przynależności do tego lub innego WWW mogą decydować różnego rodzaju czynniki. Narodowy Web obejmuje przede wszystkim materiały zarejestrowane w krajowej domenie najwyższego poziomu, takiej jak np.: .uk, .es, .cat bądź .dk. Zdarza się, że dany projekt gromadzi więcej niż jedną domenę, jak w przypadku Finlandii (oprócz domeny .fi archiwizowana jest domena autonomicznych Wysp Alandzkich – .ax)[34] czy Francji (Biblioteka Narodowa Francji gromadzi zawartość domen terytoriów zamorskich Francji)[35]. Zasoby interesujące poszczególne archiwa mogą znajdować się także poza domenami krajowymi, dlatego też stosuje się dodatkowe kryteria, które pozwalają uzupełnić je o te materiały. Zaliczyć można do nich lokalizację serwera[36], a także narodowość autora, powiązania tematyczne lub terytorialne, a także znaczenie kulturowe lub popularność wśród obywateli danego państwa[37]. Takim kryterium może być również fakt, że dane materiały kierowane są do nich, jak ma to miejsce w przypadku duńskiego Netarkivet i jest usankcjonowane tamtejszymi przepisami prawnymi. Częstą praktyką jest określanie zakresu archiwizacji w regulacjach związanych z egzemplarzem obowiązkowym, gdzie wskazywane jest, jakiego rodzaju materiały podlegają gromadzeniu[38].

W celu wypełniania swojego zadania w zakresie, jaki został przedstawiony powyżej, archiwa Webu wykorzystują odpowiednie strategie gromadzenia zasobów. Wskazać można na dwie podstawowe strategie: masową i selektywną. Pierwsza z nich polega na automatycznej archiwizacji dużych wycinków Sieci na bazie pozyskanej wcześniej listy adresów URL, druga zaś na wyborze materiałów przez osoby za to odpowiedzialne na podstawie określonych kryteriów i wskazówek[39]. Jak można zauważyć w tabeli nr 2, większość projektów łączy obie te metody, dzięki czemu możliwe jest zachowanie pełniejszego wycinka

[34] E.P. Keskitalo, *Web Archiving in Finland. Memorandum for the members of the CDNL*, 2010, s. 10, http://www.doria.fi/bitstream/handle/10024/67051/webarchivingfinland_cdnl.pdf [dostęp: 8.06.2020].

[35] S. Aubry, op.cit., s. 182.

[36] Z lokalizacji serwerów korzysta m.in. UK Web Archive, zob.: UK Web Archive, Frequently asked questions, https://www.webarchive.org.uk/en/ukwa/info/faq [dostęp: 8.06.2020].

[37] Takie kryteria wymienia m.in. archiwum Webu chorwackiego (zob.: K. Holub, I. Rudomino, op.cit., s. 3) oraz estońskiego (zob.: Eesti veebiarhiiv on Rahvusraamatukogu, Veebisaidid, https://www.nlib.ee/veebisaidid [dostęp: 8.06.2020].

[38] Np.: Act on Legal Deposit of Published Material § 2 (3), tłumaczenie ustawy nr 1439 z 22 grudnia 2004, wersja nieautoryzowana, http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html [dostęp: 8.06.2020].

[39] ISO/DTR 14873 Information and documentation — Statistics and Quality Indicators for Web Archiving, 2012, s. 9, http://netpreserve.org/resources/IIPC_project-SO_TR_14873__E__2012-10-02_DRAFT.pdf [dostęp: 8.06.2020].

Sieci. Wykorzystanie archiwizacji masowej nie jest możliwe chociażby w przypadku projektu Webarchief van Nederland ze względu na brak przepisów pozwalających Bibliotece Narodowej Holandii na takie działanie[40]. Z podobnym problemem spotkało się już wcześniej UK Web Archive[41].

Tabela 2: Zakres i strategie gromadzenia w wybranych projektach archiwizacji Webu w Europie – opracowanie własne

Nazwa projektu	Gromadzenie masowe	Gromadzenie selektywne		
		Tematyczne	Evenet harvesting	Inne
WebArchiv	domena .cz	Brak	Jest	Kolekcje tematyczne
Archives de l'internet	Domeny .fr i terytoriów zamorskich Francji oraz witryny zidentyfikowane jako część „francuskiej" Sieci	Jest	Jest	Kolekcje tematyczne
Hrvatski arhiv weba	Domena .hr	Jest	Jest	–
Vefsafn.is	Domena .is	Brak	Jest	Wybrane ważne witryny
UK Web Archive	Domena .uk, .scot, .wales, .cymru, .london, serwery zlokalizowane w Zjednoczonym Królestwie	Jest	Jest	Kolekcje tematyczne
PADICAT	Domena .cat	Jest	Jest	Kolekcje tematyczne
Netarkivet	Domena .dk oraz witryny zidentyfikowane jako część „duńskiej" Sieci	Brak	Jest	Wybrane 80–10 witryn, specjalne gromadzenie (zamykanie witryny, potrzeba badacza)
Suomalainen verkkoarkisto	Domeny .fi i .ax, serwery zlokalizowane w Finlandii	Brak	Jest	Kolekcje tematyczne, strony z wiadomościami
Ondarenet	Brak	Jest	Jest	–
Webarchief van Nederland	Brak	Brak	Jest	Kolekcje tematyczne, wybrane ważne witryny
Arquivo.pt	Domena .pt oraz witryny zidentyfikowane jako część „portugalskiej" Sieci	Brak	Jest	Kolekcja projektów naukowych i rozwojowych fundowanych przez EU

[40] Koninklijke Bibliotheek, Legal issues, https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/legal-issues [dostęp: 8.06.2020].

[41] British Library, Collection guides. UK Web Archive, https://www.bl.uk/collection-guides/uk-web-archive [dostęp 8.06.2020].

Nazwa projektu	Gromadzenie masowe	Gromadzenie selektywne		
		Tematyczne	Evenet harvesting	Inne
Archivo de la Web Española	Domena .es oraz witryny zidentyfikowane jako część „portugalskiej" Sieci	Jest	Jest	Kolekcje tematyczne i społeczności autonomicznych, witryny zagrożone usunięciem
Eesti veebiarhiivi	Najprawdopodobniej domena .ee oraz witryny zidentyfikowane jako część „estońskiej" Sieci	Jest	Jest	Kolekcje tematyczne
Web Archive	Domena .ie	Brak	Jest	Kolekcje tematyczne

Masowa archiwizacja Webu przeprowadzana jest przez analizowane w artykule inicjatywy w mocno zbliżony sposób. Zasoby nią objęte pochodzą niemal w całości z domen krajowych, jednak często bywają rozszerzane o inne zasoby (zob. Tabela 2). W większości przypadków *crawle* wykonuje się raz do roku, częściej zaś w Katalonii (2 razy)[42], Portugalii (3–4 razy)[43] oraz Danii (4 razy)[44]. *Seedlist*, czyli lista adresów URL, od których robot archiwizujący rozpoczyna cały proces, pozyskiwana jest od instytucji odpowiedzialnych za prowadzenie rejestru danej domeny[45]. W różny sposób projekty te traktują Robots Exclusion Protocol (nazywany często „*robots.txt*"), a więc skrypt zezwalający różnego rodzaju robotom internetowym na dostęp do witryny internetowej. Protokół ten jest respektowany m.in. przez archiwa Sieci w Chorwacji[46] oraz Portugalii[47], natomiast ignorowany jest chociażby przez projekty w Danii[48] i Hiszpanii[49], które powołują się na obowiązujące je przepisy o egzemplarzu obowiązkowym. Część inicjatyw wprowadza swoje wewnętrzne ograniczenia dla archiwizowanych zasobów, które wynikają przede wszystkim z przeszkód technologicznych. Jako przykład można podać Eesti Veebiarhiivi, które z tych względów rezygnuje

[42] The Web Archive of Catalonia, Mission and objectives, https://www.padicat.cat/en/about-us/what-padicat/mission-and-objectives [dostęp: 8.06.2020].

[43] Arquivo.pt, Crawling and archiving Web content, https://sobre.arquivo.pt/en/crawling-and-archiving-web-content/#qe-faq-2418 [dostęp: 8.06.2020].

[44] S.Schostag, E. Fønss-Jørgensen, op.cit., s. 110.

[45] M.in. archiwum chorwackiego Webu pozyskuje seedlist od CARNet, instytucji zarządzającej tamtejszą domena od 1993 r., zob.: K. Holub, I. Rudomino, op.cit., s. 7–8.

[46] Ibidem, s. 7.

[47] Arquivo.pt, Crawling and archiving Web content, https://sobre.arquivo.pt/en/crawling-and-archiving-web-content/#qe-faq-2407 [dostęp: 8.06.2020].

[48] Netarkivet, FAQ, http://netarkivet.dk/in-english/faq/#anchor8 [dostęp: 8.06.2020].

[49] Biblioteca Nacional de España, Technical details, http://www.bne.es/en/Colecciones/ArchivoWeb/InfoTecnica/index.html [dostęp: 8.06.2020].

z gromadzenia materiałów udostępnianych na żywo czy witryn, które wymagają zbyt dużo przestrzeni dyskowej[50].

Drugim stosowanym rozwiązaniem jest archiwizacja selektywna, która stanowi niejako uzupełnienie gromadzenia masowego, ponieważ ze względu na swój sposób działania może pozostawić luki w zbieranych zasobach. Strategia ta realizowana jest na parę sposobów, w zależności od celu w jakim została wykorzystana. Rozwiązanie to często przyjmuje postać tzw. *event harvesting*, a więc gromadzenia witryn internetowych i innych zasobów sieciowych związanych z określonym wydarzeniem. Nierzadką praktyką jest angażowanie do takich działań zewnętrznych instytucji i ekspertów, a także użytkowników archiwów[51]. Obecnie w wielu europejskich archiwach Webu możemy znaleźć kolekcje poświęcone głównie wyborom różnego szczebla (np. parlamentarnym, samorządowym, itp.)[52], a także katastrofom naturalnym[53], zamachom terrorystycznym i innym ważnym wydarzeniom[54]. Podobny charakter mają także mniejsze, ale za to szczegółowe, kolekcje tematyczne, których przykłady można znaleźć m.in. w zasobach archiwum Webu Wielkiej Brytanii[55] oraz Irlandii[56].

Innym wariantem wykorzystania strategii selektywnej jest archiwizacja tematyczna. W jej trakcie wybiera się poszczególne zasoby na podstawie wcześniej określonych kryteriów i z reguły przyporządkowuje do którejś z kilku lub kilkunastu ogólnych kategorii[57]. Takie podejście pozwala na zachowanie materiałów, które mogłyby znaleźć się poza zasięgiem archiwizacji masowej lub, ze względu na swój charakter, mogą wymagać częstszego archiwizowania. Selekcjonując zasoby Sieci w ten sposób, Hrvatski arhiv weba kieruje się dwoma rodzajami

[50] Veebisaidid, Eesti veebiarhiiv on Rahvusraamatukogu, https://www.nlib.ee/veebisaidid [dostęp: 8.06.2020].

[51] Z takiego wsparcia korzysta chociażby Biblioteka Narodowa Francji, zob.: S. Aubry, op.cit., s. 183.

[52] Np. w zbiorach katalońskiego archiwum PADICAT lub w portugalskim Arquivo.pt, zob.: The Web Archive of Catalonia, Monographics, https://www.padicat.cat/en/search-and-discover/monographics [dostęp: 8.06.2020]; Arquivo.pt, Colaborative Collections, https://sobre.arquivo.pt/en/collaborate/colaborative-collections/ [dostęp: 8.06.2020].

[53] Hrvatski arhiv weba posiada w swoich zasobach kolekcję poświęconą powodzi z 2014 r., zob.: Croatian Web Archive, Flood in Croatia, https://haw.nsk.hr/en/thematic-collections/12/flood-in-croatia-2014 [dostęp: 8.06.2020].

[54] Biblioteka Narodowa Hiszpanii zgromadziła zasoby sieciowe dotyczące zamachów terrorystycznych czy abdykacji króla Juana Carlosa, zob.: Biblioteca Nacional de España, Collections, http://www.bne.es/en/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html [dostęp: 8.06.2020].

[55] UK Web Archive, Topics and Themes, https://www.webarchive.org.uk/en/ukwa/collection [dostęp: 8.02.2020].

[56] National Library of Ireland, Web Archive Collections, https://www.nli.ie/en/udlist/web-archive-collections.aspx [dostęp: 8.06.2020].

[57] Przykłady takich zbiorów można znaleźć na witrynach Ondarenet oraz Hrvatski arhiv weba, zob.: Ondaerenet, Departamento de Cultura y Política Lingüística, http://www.ondarenet.kultura.ejgv.euskadi.eus:8085/ondarenet/ [dostęp: 8.06.2020]; Croatian Web Archive, Browse by subject, https://haw.nsk.hr/en/browse-by-subject/ [dostęp: 8.06.2020].

kryteriów: ogólnymi, którym podlegają innego rodzaju zbiory, oraz szczególnymi. Pierwsze z nich dotyczą cech łączących się z Chorwacją, takich jak narodowość autora, język, miejsce publikacji lub tematyka, natomiast do drugich zaliczają się: zawartość, autor, wydawca, struktura danych, domena, struktura i unikatowość. Na ich podstawie pracownicy Narodowej i Uniwersyteckiej Biblioteki w Zagrzebie wyszukują interesujące materiały lub rozpatrują zgłoszenia użytkowników, bibliotekarze decydują także o częstotliwości oraz innych parametrach archiwizacji[58]. Duński Netarkivet gromadzi natomiast selektywnie od 80 do 100 witryn, z czego około 60% to serwisy z wiadomościami, 30% to dynamiczne i popularne portale, a pozostałe 10% to witryny innowacyjne. Lista witryn wybieranych w ten sposób nie jest stała, każda z nich przechodzi przez powtarzaną co 6 miesięcy analizę, w wyniku której decyduje się o zasadności jej gromadzenia, głębokości archiwizacji oraz częstotliwości (w przypadku tej inicjatywy od sześciu razy dziennie do jednego razu w miesiącu)[59].

Kolejną kwestią, wymagającą omówienia po gromadzeniu archiwalnego Webu, jest jego opisywanie. W przypadku tego rodzaju zasobów ciężko doszukiwać się znanego z klasycznej archiwistyki podziału na zespoły oraz jednostki archiwalne, ponieważ poszczególne domeny (np.: „archiva.gov.pl") traktowane są raczej w sposób biblioteczny jako osobne obiekty. Mogą być one ze sobą powiązane za pomocą kolekcji, kategorii rzeczowych lub słów kluczowych. Użytkownik korzystający z materiałów pozyskiwanych masowo ma dostęp tylko do podstawowych informacji na ich temat, co podyktowane jest ich liczbą oraz wykorzystywanym oprogramowaniem. Wśród udostępnianych metadanych znajduje się adres URL, liczba wykonanych zrzutów, ich zakres chronologiczny oraz daty i odnośniki do poszczególnych kopii[60]. Zastosowanie strategii selektywnej pozwala na dołączenie dodatkowych elementów takich jak: tytuł, opis, kategoria, słowa kluczowe[61], a nawet na sporządzenie pełnego bibliotecznego opisu[62]. Swój opis posiadają także poszczególne kolekcje, w większości przypadków przybiera on dość prostą formę[63], a za dobry przykład można podać Bibliotekę Narodową Irlandii, która publikuje bogatą dokumentację na temat ich budowania[64].

---

[58] K. Holub, I. Rudomino, op.cit., s. 3–7.

[59] S. Schostag, E. Fønss-Jørgensen, op.cit., s. 111–114.

[60] Na temat informacji udostępnianych przez Wayback Machine, zob.: M. Wilkowski, *Jak korzystać w Wayback Machine,* https://wilkowski.org/waybackmachine [dostęp: 8.06.2020].

[61] Przykład z zasobu Hrvatski arhiv weba: Croatian Web Archive, Arhivi, knjižnice, muzeji, https://haw.nsk.hr/en/publikacija/1109/ [dostęp: 8.06.2020].

[62] Przykład z katalogu Biblioteki Narodowej Irlandii: National Library of Ireland's Catalogue, Dublin.ie, http://catalogue.nli.ie/Record/vtls000659084 [dostęp: 8.06.2020].

[63] Przykład z zasobu UK Web Archive: UK Web Archive, Caribbean Communities in the UK, https://www.webarchive.org.uk/en/ukwa/collection/2131 [dostęp: 8.06.2020].

[64] Zob. np.: National Library of Ireland, Remembering 1916, Recording 2016, https://www.nli.ie/GetAttachment.aspx?id=f3f10f40-6626-4692-aa51-8d7187827235 [dostęp: 8.06.2020].

Następnym zagadnieniem poddanym analizie jest udostępnianie zasobów archiwów Webu, ponieważ, jak widać w tabeli nr 1, może ono posiadać różny zakres. Znaczny wpływ mają na to przepisy dotyczące egzemplarza obowiązkowego, które często regulują kwestię udostępniania oraz związane z nim ograniczenia wynikające z praw autorskich czy ochrony danych wrażliwych. Duża część zasobów jest otwarcie udostępniana poprzez portale poszczególnych archiwów, jednak tylko cztery z nich oferują dostęp do całości, tak jak Hrvatski arhiv weba[65] lub Arquivo.pt, które udostępnia swoje zbiory rok po ich zgromadzeniu[66]. Znacznie częściej możliwe jest korzystanie z materiałów pozyskanych strategią selektywną, ponieważ w trakcie tego procesu często wykorzystuje się zgody właścicieli witryn na wykonanie i późniejsze wykorzystanie ich kopii. Na takich zasadach możliwy jest dostęp do kolekcji tematycznych UK Web Archive, zaś pozostała część zasobów dostępna jest na terenie bibliotek z prawem do egzemplarza obowiązkowego[67]. Zdarzają się również przypadki ograniczenia korzystania ze zgromadzonych zasobów do przeglądania ich wyłącznie na odpowiednich stanowiskach, jak ma to miejsce np. we Francji[68], Holandii[69] oraz Danii, gdzie funkcjonują rygorystyczne przepisy związane z jego udostępnianiem[70].

Zasoby archiwów Sieci udostępniane są za pomocą specjalnie przygotowanych serwisów dostępnych publicznie w Internecie lub na odpowiednich stanowiskach w czytelniach bibliotek. Ich przeszukiwanie odbywa się za pomocą adresów URL, słów kluczowych, a także poprzez kolekcje oraz kategorie tematyczne, które mogą dzielić się na bardziej szczegółowe podkategorie[71]. Stopniowo wprowadzane są bardziej zaawansowane metody, wykorzystujące np. wyszukiwanie pełnotekstowe. Przywołać można tu silnik wyszukiwawczy SHINE przygotowany na potrzeby UK Web Archive i projektu BUDDAH, który pozwala na stosowanie zaawansowanych filtrów do wyszukiwania, a także przeprowadzenie analizy trendów[72].

---

[65] Croatian Web Archive, For publishers, https://haw.nsk.hr/en/for-publishers/ [dostęp: 8.06.2020].

[66] Arquivo.pt, Access to archived content, https://sobre.arquivo.pt/en/help/access-to-archived-contents/ [dostęp: 8.06.2020].

[67] British Library, Collection guides. UK Web Archive, https://www.bl.uk/collection-guides/uk-web-archive [dostęp: 08.06.2020].

[68] Bibliothèque nationale de France, Archives de l'internet, https://www.bnf.fr/fr/archives-de-linternet [dostęp: 8.06.2020].

[69] Koninklijke Bibliotheek, Web archiving, https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving [dostęp: 27.11.2018].

[70] S. Schostag, E. Fønss-Jørgensen, op.cit., s. 17.

[71] Z takich rozwiązań korzysta m.in. Biblioteka Narodowa Francji, zob.: S. Aubry, op.cit., s. 185–188.

[72] UK Web Archive, About SHINE, https://www.webarchive.org.uk/shine [dostęp: 8.06.2020].

Tabela 3: Oprogramowanie stosowane w wybranych projektach archiwizacji Webu w Europie – opracowane na podstawie Wikipedia: List of Web archiving initiatives, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 7.06.2020].

Nazwa projektu	Gromadzenie	Udostępnianie	Inne
WebArchiv	Heritrix	Wayback	Seeder
Archives de l'internet	Heritrix	Wayback	BCWeb, NetarchiveSuite
Hrvatski arhiv weba	Heritrix, DAMP	Wayback, Lucene	–
Vefsafn.is	Heritrix	OpenWaybaack	–
UK Web Archive	Heritrix	Wayback	Solr, Web Curator Tool
PADICAT	Heitrix	Wayback, Wera	Web Curator Tool, Curator Archiving Tool, NutchWAX
Netarkivet	Heitrix	Wayback	NetarchiveSuite, Solr,
Suomalainen verkkoarkisto	Heritrix	Wayback	Solr
Ondarenet	Heritrix	Wayback	Web Curator Tool, NutchWAX
Webarchief van Nederland	Heritrix	Wayback	Web Curator Tool, KB e-Depot System
Arquivo.pt	Heritrix, Brozzler	Wayback	Solr, NutchWAX, pywb, własne specjalistyczne oprogramowanie
Archivo de la Web Española	Heritrix	OpenWayback	Solr
Eesti veebiarhiivi	Heritrix, Squidwarc	Wayback	Custom Curator Tool, pywb
Web Archive	–	–	Archive-It

Analizowane archiwa wykorzystują w swojej pracy praktycznie jednolite oprogramowanie pozwalające na gromadzenie i udostępnianie zasobów Webu (zob. Tabela 3). Powszechnie używany jest *crawler* Heritrix. Jest to stworzony przez Internet Archive robot internetowy przeszukujący Sieć i pobierający kopie materiałów do zbiorów archiwum[73]. Do korzystania z dawnego WWW wykorzystuje się technologię OpenWayback opracowaną przez tę samą fundację i obecnie rozwijaną przez IIPC[74]. Oba te narzędzia są otwarcie udostępniane i możliwe jest ich przystosowanie do własnych potrzeb. Poszczególne archiwa korzystają też z innych programów przeznaczonych dla tego rodzaju inicjatyw. Wskazać można przede wszystkim na narzędzia służące zarządzaniu tego rodzaju pro-

[73] GitHub, Heritrix wiki, https://github.com/internetarchive/heritrix3/wiki [dostęp: 8.06.2020].
[74] GitHub, OpenWayback wiki, https://github.com/iipc/openwayback/wiki [dostęp: 8.06.2020].

jektami (jak np. Web Curator Tool[75] lub NetarchiveSuite[76]), które umożliwiają między innymi planowanie *crawli*, wskazywanie witryn do archiwizacji lub sprawdzanie jakości wykonanych *snapshotów*. Inną grupę software'u stosowaną przez analizowanie archiwa są silniki wyszukiwawcze, w ich przypadku są to NutchWAX[77] oraz Solr[78]. Należy zwrócić uwagę na jeden wyjątek – Biblioteka Narodowa Irlandii korzysta obecnie z usług Archive-It i w tym serwisie dostępne są stworzone przez nią kolekcja witryn WWW[79].

Podsumowanie

Jak można zauważyć z wyników przeprowadzonej analizy europejskich projektów archiwizacji Webu, stosują one bardzo zbliżone praktyki. Przyczyniać się do tego może fakt, iż ich organizatorami są prawie wyłącznie biblioteki narodowe oraz większość z nich należy do IIPC. Instytucja ta poszukuje dobrych praktyk oraz rozwija oprogramowanie dla tego rodzaju działalności, a wspiera ją w tym fundacja Internet Archive, która tym sposobem implementuje zaproponowane rozwiązania. Europejskie archiwa Sieci gromadzą jej narodowy wycinek, za który najczęściej uznaje się krajową domenę najwyższego poziomu, a także materiały znajdujące się poza nimi, jeżeli są powiązane z danym państwem lub adresowane do jego obywateli. W celu jak najpełniejszej archiwizacji łączy się dwie strategie – masową i selektywną (tematyczną oraz *event harvesting*). Ze względu na rozmiary tego rodzaju zbiorów, rzadko kiedy posiadają one bogaty opis informacyjny. Zagadnienie opisu i metadanych może wymagać w przyszłości większej uwagi ze względu na późniejsze wykorzystanie archiwalnego Webu. Z tej perspektywy problemem jest również, ograniczony w wielu przypadkach, dostęp do części lub całości zbiorów, co wynika m.in. z prawa autorskiego oraz przepisów dotyczących egzemplarza obowiązkowego lub danych osobowych.

Na zakończenie warto dodać, iż kwestia dostępu do zasobów dawnego WWW oraz jego wykorzystanie w nauce, ale nie tylko, powinna stać się priorytetem w rozwoju archiwistyki Sieci. Gromadzenie latami setek terabajtów danych bez możliwości ich ponownego użycia wydaje się być bezzasadne, zwłaszcza gdy zaczynają się rozwijać takie kierunki badań jak *web studies* lub *web history*. Tego rodzaju badania, w których wykorzystywano zbiory analizowanych archiwów, były już przeprowadzane. Jako przykład przytoczyć można dwie analizy,

[75] Web Curator Tool Documentation, Read the Docs, https://webcuratortool.readthedocs.io/en/latest/ [dostęp: 9.06.2020].

[76] SBForge, NetarchiveSuite, https://sbforge.org/display/NAS/NetarchiveSuite [dostęp: 9.06.2020].

[77] SourceForge, NutchWAX, http://archive-access.sourceforge.net/projects/nutchwax/index.html [dostęp: 9.06.2020].

[78] Apache Lucene, Apache Solr, https://lucene.apache.org/solr/ [dostęp: 9.06.2020].

[79] Profil National Library of Ireland: Archive-It, National Library of Ireland, https://archive-it.org/home/nli [dostęp: 9.06.2020].

opublikowane w przywoływanej już książce *The Web as the History,* dotyczące Sieci brytyjskiej oraz duńskiej. W pierwszym przypadku badacze zainteresowali się rozwojem poszczególnych brytyjskich subdomen, ze szczególnym uwzględnieniem subdomeny akademickiej – .ac.uk[80]. Drugie z przywołanych badań dotyczyło nazw duńskich domen i ich wykorzystania do badania tamtejszego Webu[81].

Bibliografia

A Research Infrastructure for the Study of Archived Web Materials, About RESAW, http://resaw.eu/about/ [dostęp: 8.06.2020].

A Research Infrastructure for the Study of Archived Web Materials, Participants, http://resaw.eu/participants/ [dostęp: 8.06.2020].

Act on Legal Deposit of Published Material § 2(3), tłumaczenie ustawy nr 1439 z 22 grudnia 2004, wersja nieautoryzowana, http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html [dostęp: 8.06.2020].

AlSum A., Weigle M.C., Nelson M.L., Sompel H. Van de, *Profiling web archive coverage for top-level domain and content language*, „International Journal on Digital Libraries" 2014, t. 14, nr 3–4, s. 149, https://link.springer.com/article/10.1007%2Fs00799-014-0118-y [dostęp: 7.06.2020].

Apache Lucene, Apache Solr, https://lucene.apache.org/solr/ [dostęp: 9.06.2020].

Archive-It, Archive-It Blog – About us, https://archive-it.org/blog/learn-more/ [dostęp: 7.06.2020].

Archive-It, National Library of Ireland, https://archive-it.org/home/nli [dostęp: 9.06.2020].

Arquivo.pt, Access to archived content, https://sobre.arquivo.pt/en/help/access-to-archived-contents/ [dostęp: 8.06.2020].

Arquivo.pt, Colaborative Collections, https://sobre.arquivo.pt/en/collaborate/colaborative-collections/ [dostęp: 8.06.2020].

Arquivo.pt, Crawling and archiving Web content, https://sobre.arquivo.pt/en/crawling-and-archiving-web-content/ [dostęp: 8.06.2020].

Aubry S., *Web Archives as a New Library Service: the Experience of the National Library of France*, „LIBER Quarterly" 2010, t. 20, nr 2, s. 179–199, https://www.liberquarterly.eu/articles/10.18352/lq.7987/ [dostęp: 7.06.2020].

Biblioteca Nacional de España, Collections, http://www.bne.es/en/Colecciones/ArchivoWeb/Subcolecciones/selectivas.html [dostęp: 8.06.2020].

Biblioteca Nacional de España, History of the collection, http://www.bne.es/en/Colecciones/ArchivoWeb/Historia/index.html [dostęp: 7.06.2020].

[80] E. T. Meyer, T. Yasseri, S. A. Hale, J. Cowls, R. Schroeder, H. Margetts, *Analysing the UK web domain and exploring 15 years of UK universities on the web*, [w:] *The Web as History*, s. 23–44, https://www.jstor.org/stable/j.ctt1mtz55k.7?seq=1#metadata_info_tab_contents [dostęp: 7.06.2020].

[81] N. Brügger, D. Laursen and J. Nielsen, *Exploring the domain names of the Danish web*, [w:] *The Web as History*, s. 62–80, https://www.jstor.org/stable/j.ctt1mtz55k.9?seq=1#metadata_info_tab_contents [dostęp: 7.06.2020].

Biblioteca Nacional de España, Technical details, http://www.bne.es/en/Colecciones/ArchivoWeb/InfoTecnica/index.html [dostęp: 8.06.2020].

Bibliothèque nationale de France, Archives de l'internet, https://www.bnf.fr/fr/archives-de-linternet [dostęp: 8.06.2020].

Big UK Domain Data for the Arts and Humanities, Aims and objectives, https://buddah.projects.history.ac.uk/about/aims-and-objectives/ [dostęp: 8.06.2020].

British Library, Collection guides. UK Web Archive, https://www.bl.uk/collection-guides/uk-web-archive [dostęp: 8.06.2020].

Costea M.D., *Report on the Scholarly Use of Web Archives*, Aarhus 2018, http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf [dostęp: 7.06.2020].

Croatian Web Archive, Arhivi, knjižnice, muzeji, https://haw.nsk.hr/en/publikacija/1109/ [dostęp: 8.06.2020].

Croatian Web Archive, Browse by subject, https://haw.nsk.hr/en/browse-by-subject/ [dostęp: 8.06.2020].

Croatian Web Archive, Flood in Croatia, https://haw.nsk.hr/en/thematic-collections/12/flood-in-croatia-2014 [dostęp: 8.06.2020].

Croatian Web Archive, For publishers, https://haw.nsk.hr/en/for-publishers/ [dostęp: 8.06.2020].

Departamento de Cultura y Política Lingüística, Ondaerenet,http://www.ondarenet.kultura.ejgv.euskadi.eus:8085/ondarenet/ [dostęp: 8.06.2020].

Eesti Rahvusraamatukogu, Veebisaidid, https://www.nlib.ee/veebisaidid [dostęp: 8.06.2020].

Farag M.M., Lee S., Fox E.A., *Focused crawler for events*, „International Journal on Digital Libraries" 2018, t. 19, nr. 1, s. 3–19, https://link.springer.com/article/10.1007/s00799-016-0207-1 [dostęp: 7.06.2020].

Geereart F., Soyez S., The first steps towards a Belgian web archive: a federal strategy (materiały z konferencji IIPC Web Archiving Conference 2019, Zagrzeb, 6–7 czerwca 2019), http://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-FRIEDEL_GEERAERT__SEBASTIEN_SOYEZ-The_first_steps_towards_a_Belgian_web_archive-a_federal_strategy.pdf [dostęp: 9.06.2020).

GitHub, Heritrix wiki, https://github.com/internetarchive/heritrix3/wiki [dostęp: 8.06.2020].

GitHub, OpenWayback wiki, https://github.com/iipc/openwayback/wiki [dostęp: 8.06.2020].

Gomez D., Miranda J., Costa M., *A survey on web archiving initiatives*, [w:] *Research and Advanced Technology for Digital Libraries. International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26–28, 2011. Proceedings*, oprac. i red. S. Gradmann, F. Borri, C. Meghini, H. Schuldt, Berlin 2011, s. 410–413, https://link.springer.com/chapter/10.1007/978-3-642-24469-8_41 [dostęp: 7.06.2020].

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), s. 1–12, http://library.ifla.org/1092/1/090-holub-en.pdf [dostęp: 7.06.2020].

International Internet Preservation Consortium, About IIPC, http://netpreserve.org/about-us/ [dostęp: 8.06.2020].

International Internet Preservation Consortium, IIPC members, http://netpreserve.org/about-us/members/ [dostęp: 8.06.2020].

ISO/DTR 14873 Information and documentation — Statistics and Quality Indicators for Web Archiving, 2012, s. 9, http://netpreserve.org/resources/IIPC_project-SO_TR_14873__E__2012-10-02_DRAFT.pdf [dostęp: 8.06.2020].

Keskitalo E.P., *Web Archiving in Finland. Memorandum for the members of the CDNL*, 2010, s. 10, http://www.doria.fi/bitstream/handle/10024/67051/webarchivingfinland_cdnl.pdf [dostęp: 8.06.2020].

Kłębczyk F., *Archiwizacja zasobów Internetu – kierunki i wyzwania*, „Archiwista Polski" 2012, nr 3(67), s. 105–112.

Koninklijke Bibliotheek, Legal issues, https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/legal-issues [dostęp: 8.06.2020].

Koninklijke Bibliotheek, Web archiving, https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving [dostęp: 27.11.2018].

Laboratorium Cyfrowe Humanistyki UW, Toruńskie Konfrontacje Archiwalne i problemy archiwizacji Webu, https://lach.edu.pl/blog/2017/12/11/torunskie-konfrontacje-archiwalne-problemy-archiwizacji-webu/ [dostęp: 7.06.2020].

Library of Congress, Digital Collections, https://www.loc.gov/collections/?fa=original-format:archived+web+site [dostęp: 7.06.2020].

National Library of Ireland, Irish Domain Web Archive, https://www.nli.ie/en/irish-domain-web-archive.aspx [dostęp: 8.06.2020].

National Library of Ireland, Remembering 1916, Recording 2016, https://www.nli.ie/GetAttachment.aspx?id=f3f10f40-6626-4692-aa51-8d7187827235 [dostęp: 8.06.2020].

National Library of Ireland, Web Archive Collections, https://www.nli.ie/en/udlist/web-archive-collections.aspx [dostęp 8.06.2020].

National Library of Ireland's catalogue, Dublin.ie, http://catalogue.nli.ie/Record/vtls000659084 [dostęp: 8.06.2020].

Netarkivet, FAQ, http://netarkivet.dk/in-english/faq/#anchor8 [dostęp: 8.06.2020].

NetLab, Mission, http://www.netlab.dk/netlab/mission/ [dostęp: 8.06.2020].

Rosa A., *Human trace on the Internet – the issue of archiving the Web from the point of view of anthropology-oriented archival science*, „Archiwa – Kancelarie – Zbiory" 2015, t. 6(8), s. 193–205, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2015.006 [dostęp: 7.06.2020].

SBForge, NetarchiveSuite, https://sbforge.org/display/NAS/NetarchiveSuite [dostęp: 9.06.2020].

Schostag S., Fønss-Jørgensen E., *Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective*, „Microform & Digitization Review" 2012, t. 41, nr 3–4, s. 110–120, https://www.degruyter.com/view/journals/mfir/41/3-4/article-p110.xml [dostęp: 7.06.2020].

Sobczak A., *Internet jako globalne archiwum społeczne – rozważania na temat roli Internetu w dokumentowaniu dziejów ludzkości*, [w:] *Nowa archiwistyka – archiwa i archiwistyka w poźnonowoczesnym kontekście kulturowym*, Toruńskie Konfrontacje Archiwalne, t. 4, oprac. i red. W. Chorążyczewski, W. Piasek, A. Rosa, Toruń 2014, s. 237–247.

SourceForge, NutchWAX, http://archive-access.sourceforge.net/projects/nutchwax/index.html [dostęp: 9.06.2020].

The SAGE Handbook of Web History, oprac. i red. N. Brügger, I. Milligan, Thousand Oaks 2018.

The Web Archive of Catalonia, Mission and objectives, https://www.padicat.cat/en/about-us/what-padicat/mission-and-objectives [dostęp: 8.06.2020].

The Web Archive of Catalonia, Monographics, https://www.padicat.cat/en/search-and-discover/monographics [dostęp: 8.06.2020].

The Web as History: Using Web Archives to Understand the Past and the Present, oprac. i red. N. Brügger, R. Schroeder, Londyn 2017, https://www.jstor.org/stable/j.ctt1mtz55k [dostęp: 7.06.2020].

UK Web Archive, About SHINE, https://www.webarchive.org.uk/shine [dostęp: 8.06.2020].

UK Web Archive, Caribbean Communities in the UK, https://www.webarchive.org.uk/en/ukwa/collection/2131 [dostęp: 8.06.2020].

UK Web Archive, Frequently asked questions, https://www.webarchive.org.uk/en/ukwa/info/faq [dostęp: 8.06.2020].

UK Web Archive, Topics and Themes, https://www.webarchive.org.uk/en/ukwa/collection [dostęp: 8.02.2020].

Vernalte F.P., Maciá S.M., Capturing the Basque Web (materiały z konferencji LIDA 2009, Dubrownik i Zadar (Chorwacja), 25–29 maja 2009), s. 8–9, http://eprints.rclis.org/13164/1/EN_Lida_paper_Ondarenet_APA.pdf [dostęp: 25.11.2018].

Web Archiving, oprac. i red. J. Masanès, Berlin–Heidelberg 2006.

Web Curator Tool Documentation, Read the Docs, https://webcuratortool.readthedocs.io/en/latest/ [dostęp: 9.06.2020].

webArch CKC UW, Pracownia archiwizacji Webu CKC UW, https://webarch.uw.edu.pl/pracownia/ [dostęp: 8.06.2020].

Wikipedia, Internet Memory Foundation, https://en.wikipedia.org/wiki/Internet_Memory_Foundation [dostęp: 8.06.2020].

Wikipedia, List of Web archiving initiatives, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives [dostęp: 7.06.2020].

Wikipedia, World Wide Web a Internet, https://pl.wikipedia.org/wiki/World_Wide_Web#World_Wide_Web_a_Internet [dostęp: 7.06.2020].

Wilkowski M., *Jak korzystać w Wayback Machine*, https://wilkowski.org/waybackmachine [dostęp: 8.06.2020].

Wilkowski M., *Oddolne archiwizacje Internetu jako działania społeczne*, „Archiwa – Kancelarie – Zbiory" 2015, t. 6(8), s. 207–220, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2015.007 [dostęp: 7.06.2020].

Woźniak W., *Archiwizacja Internetu – próba podsumowania dotychczasowych prac i ustaleń*, „Archiwa – Kancelarie – Zbiory" 2015, t. 10(12), s. 75–98, https://apcz.umk.pl/czasopisma/index.php/AKZ/article/view/AKZ.2019.004 [dostęp: 7.06.2020].