1. Archiwizacja webu
	1. Czym jest archiwizacja webu?

2.1.1 Archiwizacja webu to proces pobierania z World Wide Web stron internetowych i zawartych w nich informacji oraz przechowywania ich w archiwum. Archiwizacja webu jest procesem podobnym do tradycyjnej archiwizacji dokumentów papierowych lub drukowanych; informacje są wybierane, przechowywane, zabezpieczane i udostępniane użytkownikom. Zarchiwizowane strony internetowe zazwyczaj są dostępne do wykorzystania przez rząd, przedsiębiorstwa, organizacje, naukowców, historyków i społeczeństwo[[1]](#footnote-1). Podobnie jak w tradycyjnych archiwach, strony webowe są gromadzone i przechowywane przez archiwistów zwanych „archiwistami webu”.

* + 1. Ponieważ Internet zawiera ogromną liczbę stron internetowych i informacji, archiwiści webowi zazwyczaj wykorzystują zautomatyzowane procesy zbierania stron internetowych. Procesy te polegają na *harvestowaniu* (tj. masowym pobieraniu adresów) stron internetowych z ich lokalizacji w Internecie za pomocą specjalnie zaprojektowanego oprogramowania. Ten typ oprogramowania, znany jako "robot indeksujący", przeszukuje zasoby Internetu, kopiując i zapisując informacje w trakcie pobierania stron internetowych. Zarchiwizowane strony internetowe i zawarte w nich informacje są udostępniane online w ramach kolekcji archiwów internetowych. Możemy je przeglądać, czytać i nawigować w nich tak, jak w sieci internetowej, jednak są one zachowywane jedynie w formie "migawek" informacji w określonych punktach czasowych.
		2. Niektóre organizacje używają prostych narzędzi i procesów do archiwizowania treści własnych stron WWW. Biblioteki narodowe, archiwa krajowe oraz różne grupy są również zaangażowane w archiwizowanie w sposób szczegółowy treści o znaczeniu kulturowym. Inną grupę stanowią organizacje archiwizujące własne treści internetowe w celach biznesowych, prawnych bądź regulacyjnych, wykorzystujące komercyjne oprogramowanie i usługi archiwizacji webu[[2]](#footnote-2). Największą organizacją archiwizującą web jest Internet Archive, której celem jest utrzymanie archiwum całej sieci WWW.
	1. Typy archiwizacji webu

2.2.1 Istnieją 3 główne techniczne metody archiwizacji treści internetowych: archiwizacja strony po stronie klienta, archiwizacja sieci w oparciu o transakcje oraz archiwizacja strony po stronie serwera[[3]](#footnote-3). Archiwizacja po stronie klienta to najpopularniejsza metoda, która może być przeprowadzana zdalnie i na dużą skalę. Podejścia oparte na transakcjach i realizowane po stronie serwera wymagają aktywnej współpracy z właścicielami serwerów i muszą być wdrażane indywidualnie dla każdego przypadku.

* + 1. Uwaga: Wszystkie 3 podejścia różnią się od tworzenia kopii zapasowej strony internetowej, która pozwala jedynie na przywrócenie strony z zapisanych plików w razie problemów. Opisane powyżej metody dotyczą archiwizacji stron internetowych, co oznacza, że strony mogą być gromadzone, przechowywane, dostępne dla użytkowników z opcją nawigowania po nich w sposób podobny do oryginalnej witryny „na żywo”.
		Więcej szczegółów na temat typów archiwizacji webu można znaleźć w Dodatku A.
	1. Po co archiwizować strony internetowe?

2.3.1 Wiele organizacji tworzy strony internetowe w ramach komunikacji ze społeczeństwem i innymi organizacjami, strony internetowe są bowiem potężnym narzędziem wymiany informacji. Dokumentują one publiczny charakter organizacji i ich interakcje z odbiorcami i klientami. Ponadto strony internetowe są coraz częściej jednym miejscem publikowania i udostępniania informacji. Z tego powodu stają się kluczową częścią dokumentacji i tożsamości organizacji bądź jednostki.

* + 1. Internet zapewnia dostęp do aktualnych informacji dzięki regularnemu aktualizowaniu i stałemu rozwojowi stron internetowych. Jest to jedna z największych zalet sieci, jednak oznacza także, że informacje dostarczane w ten sposób mogą czasami być postrzegane jako efemeryczne z natury i mają małą lub żadną trwałą wartość. Powoduje to niebezpieczeństwo, że można je łatwo utracić, zanim zostaną zachowane jako dowód do celów biznesowych lub historycznych.

2.3.3 Znaczna część sieci internetowej rozwijanej w początkowych latach Internetu oraz informacje, które kiedyś w niej były, zniknęła na zawsze; z wczesnych treści zamieszczanych online od początku lat dziewięćdziesiątych do około 1997 r., przetrwało bardzo mało informacji webowych. Tak było przed rozpoznaniem trwałej wartości starszych informacji publikowanych w Internecie do chwili podjęcia pierwszych działań związanych z archiwizacją sieci, które rozpoczęły się w 1996 roku[[4]](#footnote-4). Od lat dziewięćdziesiątych XX wieku sieć WWW obok znaczenia kulturowego, zaczęła stawać się coraz ważniejszym źródłem informacji. W rezultacie stała się integralną częścią innych działań, takich jak badania, odwołania i cytowania. W działalności, która dotychczas opierała się na fizycznych dokumentach, coraz częściej korzysta się ze stron WWW i zamieszcza linki do dokumentów przechowywanych na stronach internetowych. Archiwizacja w Internecie jest ważnym procesem, który zapewnia ludziom i organizacjom możliwość uzyskania długoterminowego dostępu do wiedzy i ponownego jej wykorzystania, a także odzyskiwania informacji.

* + 1. Archiwizacja webu może być względnie tanim i wydajnym procesem, zależy to jednak od zastosowanych metod. W idealnej sytuacji archiwa sieciowe zbierają zasoby w oryginalnej formie i mogą je udostępniać tak, jak to wyglądało „na żywo”, dostarczając zapis treści internetowych z określonego dnia i określonej godziny. Po zarchiwizowaniu strony internetowej zachowywany jest kontekst zawartej w niej informacji, co oznacza, że użytkownicy mogą wyświetlać informacje w kontekście, w którym zostały one pierwotnie przedstawione.
		2. Kopie zapasowe stron internetowych nie zawsze działają jak archiwa internetowe – szczególnie gdy strony internetowe używają aktywnych skryptów. Kopie zapasowe tego typu stron zawierają tylko kod programowania, nie zostają zebrane z Internetu i opatrzone znacznikiem czasu (znacznik czasu to odczytywana komputerowo data i godzina, którą robot indeksujący dokłada do każdego pliku, który zbiera. Dzięki temu zarchiwizowana strona internetowa jest realną reprezentacją strony internetowej w momencie jej archiwizacji). W przypadku stron internetowych, które używają tylko płaskiego kodu HTML oraz do osobistej archiwizacji własnej witryny internetowej, kopie zapasowe są akceptowalne, o ile zawierają daty utworzenia i daty zmian w plikach zapasowych.
		3. Archiwizacja stron internetowych daje organizacjom szansę zapewnienia dostępu do starszych informacji, które niekoniecznie chciałyby zachować na swojej "żywej" stronie WWW. Ewidencja powstała z inicjatywy Web Continuity w The National Archives wykazuje znaczący i ciągły popyt ze strony użytkowników na dostęp do starszych treści, które organizacja może uznać za nieaktualne lub nieważne. Archiwizacja i zapewnienie dostępu do tych treści staje się częścią szerszych działań związanych z zarządzaniem informacjami i dokumentami. W związku z tym archiwizacja w Internecie może przyczynić się do pozytywnego wizerunku organizacji jako zdolnej do skutecznego zarządzania swoimi zasobami informacyjnymi.
1. Zarządzanie informacją i dokumentami

3.1 Strony internetowe jako dokumenty

3.1.1 Zarządzanie informacją na stronie internetowej powinno być częścią szerszego podejścia do zarządzania informacją i dokumentacją. Informacja na stronach WWW powinna podlegać tym samym zasadom zarządzania, przeglądania i selekcji, co zarządzanie wszelkimi innymi dokumentami tworzonymi przez organizację.

* + 1. Dokumenty i informacje muszą być zarządzane i przechowywane z różnych powodów i służyć do różnych celów: do bieżącego wykorzystania biznesowego, do celów prawnych, jako dowody, a także w celach historycznych i kulturowych. Podobnie jak dokumenty papierowe i pliki cyfrowe, witryny internetowe wspierają bieżącą i przyszłą działalność użytkownika i jego organizacji. Jeśli strony internetowe i zawarte na nich informacje są traktowane jak dokumenty, to przechwytywanie, zarządzanie i pobieranie tych informacji tak długo, jak jest to potrzebne, stanowi potężny i pozytywny wkład w zarządzanie wszystkimi niezbędnymi dokumentami i informacjami.
		2. Dla nowicjuszy w zarządzaniu dokumentami, podstawowe wskazówki znajdują się na stronie brytyjskiego archiwum rządowego The National Archives: www.nationalarchives.gov.uk/information-management/projects-andwork/records-management-code.htm. Informacje na niej zawarte to część szerszego zestawu wskazówek dla zarządzających dokumentami. Są one przeznaczone dla organizacji rządowych i sektora publicznego, jednak zawierają ogólne zasady, użyteczne w zarządzaniu dokumentacją i informacją w szerszym kontekście.
		3. Należy pamiętać, że kopie zapasowe witryny internetowej nie są przeznaczone do używania jako archiwa internetowe, są one natomiast przydatne do bieżących celów biznesowych. Trzeba wziąć pod uwagę, że odtworzenie strony z zapisanych plików może nie być możliwe bez szczególnej pracy menedżera sieci lub projektanta.
		4. Należy zawsze kilkakrotnie sprawdzić zawartość strony przed jej opublikowaniem w sieci. Nawet jeśli dana informacja jest dostępna w Internecie tylko przez jeden dzień, może zostać zarchiwizowana lub zbuforowana przez wyszukiwarki.
		5. Jak wspomniano w sekcji 2, strony internetowe zmieniają się i ewoluują w czasie w miarę ich aktualizacji. Należy dokumentować wygląd strony internetowej oraz informacje na niej zawarte w kolejnych momentach czasu jako wartościowy zapis, który może okazać się niezbędny w celach dowodowych lub prawnych.
	1. Wybór i gromadzenie

3.2.1 Strony internetowe są same w sobie zapisem tego, jak organizacja chciała zaprezentować się publicznie i jakie informacje przekazała. Strony WWW mogą również zawierać dokumenty, takie jak protokoły z zebrań, raporty, strategie i plany. Wszystkie informacje i dokumenty na stronach internetowych są zapisami działań. Mają wartość jako aktywa dla ludzi i organizacji, które je stworzyły, ponieważ czas i pieniądze zostały zainwestowane w ich tworzenie i zarządzanie.

* + 1. Należy rozważyć wartość strony WWW i jej zawartość. Czy zawiera treści o wartości biznesowej lub historycznej? Czy ta zawartość jest przechowywana lub zachowana w innym miejscu, na przykład na współdzielonym dysku sieciowym lub w systemie zarządzania dokumentami elektronicznymi? Można korzystać z zasad dotyczących wartości biznesowej, dokumentacyjnej i historycznej przy ocenie informacji i dokumentów umieszczonych na stronie WWW oraz ich związku z innymi dokumentami, które należy zachować. Na tej podstawie trzeba zdecydować, które informacje i jak długo należy przechowywać. Na przykład dokumenty finansowe są zwykle przechowywane przez co najmniej 7 lat.
		2. Ogólnie rzecz biorąc, należy dążyć do zachowania neutralności pod względem formatu danych. Oznacza to, że zarządzanie dokumentami i informacjami nie powinno zależeć od formatu plików (np. arkuszy kalkulacyjnych, dokumentów PDF, obrazów, stron internetowych itd.), lecz odbywać się zgodnie z ich przeznaczeniem [na etapie ich tworzenia – *przyp. tłum.*] oraz ich dotychczasowym wykorzystaniem.
		3. Należy zastanowić się, jak często trzeba wykonać archiwizację strony, na przykład raz w roku, dwa razy w roku, co trzy miesiące, a może tylko jeden raz. Zależy to od częstości zmian dokonywanych na stronie, a także od względnej ważności treści. Strona internetowa może wymagać częstszego archiwizowania w określonych momentach, na przykład, jeśli pojawia się szczególnie ważne wydarzenie, które wymusza, by strona internetowa zmieniała się regularnie.
		4. Inne (nietechniczne) względy związane z archiwizacją stron internetowych dotyczą zakresu i skali gromadzenia [dokumentów i/lub informacji – *przyp. tłum.*]. Ich rozważenie może również pomóc zdecydować, które rozwiązanie techniczne zastosować, w zależności od tego, co należy zachować i dlaczego. Strony internetowe rządu centralnego są selekcjonowane zgodnie z operacyjną polityką wyboru (Operational Selection Policy 27), która jest dostępna online: nationalarchives.gov.uk/documents/information-management/osp27.pdf.
		5. Jeśli witryna i jej treść są uważane za wartościowe, należy zorganizować ich archiwizację. Więcej informacji na ten temat znajduje się w rozdziale 5 pt. „Archiwizacja witryny: co możesz zrobić”. Należy pamiętać, że nie wszystkie treści na stronie internetowej można przechwycić poprzez archiwizację webu.
	1. Wskazówki i porady dotyczące archiwizacji i stron internetowych

3.3.1 Zapisywanie i archiwizacja

* Nie należy zapisywać wszystkich treści internetowych na nośnikach wymiennych, takich jak płyty CD lub DVD - są one podatne na utratę danych i fizyczną degradację w czasie, zwłaszcza jeśli w posiadaniu jest tylko jedna ich kopia.
* Podczas tworzenia kopii zapasowych witryny w celach biznesowych lepszym sposobem jest zapisanie ich na dysku sieciowym, który z kolei jest kopiowany do innego źródła danych.
* Jeśli jedyną opcją jest zapisanie całej zawartości sieci na dysku CD lub DVD, należy wykonać kilka ich kopii i przynajmniej jedną z nich przechowywać w odrębnym miejscu. Co 6 miesięcy należy sprawdzać działanie wszystkich kopii, a co ok. 3 lata kopiować ich zawartość na nowe dyski CD lub DVD. Takie postępowanie pomaga zapobiec stratom, które mogłyby wyniknąć z uszkodzenia, degradacji w czasie, bądź kradzieży fizycznego nośnika.
	+ 1. Strony internetowe a archiwizacja

Należy rozważyć kilka spraw ułatwiających archiwizowanie stron internetowych. Bliższe informacje zamieszczono poniżej. Przede wszystkim warto omówić zawartość archiwizowanych elementów z menedżerem strony lub projektantem.

* Jeśli to możliwe, dobrze jest przechowywać całą zawartość pod jednym głównym adresem URL (skrót URL oznacza *Uniform Resource Locator* - globalny adres zasobów i dokumentów w sieci). Dzięki temu cała zawartość strony internetowej może być łatwo zindeksowana i zarchiwizowana. To ułatwia identyfikację całej przechwyconej witryny.
* Najlepiej używać "przyjaznych" adresów URL. Tworzenie czytelnych dla człowieka lub „przyjaznych” adresów URL to dobra praktyka z wielu powodów (np. zwiększenie użyteczności, bezpieczeństwa i optymalizacja pod kątem działania wyszukiwarek).
* W przypadku używania skryptów (takich jak JavaScript) na swojej witrynie, należy dodać alternatywnie zwykły HTML, co poprawi dostępność zawartości dla użytkowników i wspiera archiwizację. W miarę możliwości warto dodawać statyczne linki lub "podstawowe kotwice strony" zamiast dynamicznie generowanych adresów URL. Jeśli strona zawiera treści multimedialne i strumieniowane, należy dodać do nich alternatywne rozwiązanie, na przykład pobieranie progresywne wraz z przesyłaną treścią.
* Należy też sprawdzać stronę WWW pod kątem dostępności zgodnie ze standardem opracowanym przez W3C. Strona internetowa, która została zaprojektowana jako W3C Web Accessible, powinna również być łatwa do zarchiwizowania[[5]](#footnote-5).
* Należy udostępniać mapy witryny w XML, które zawierają listy i linki do całej zawartości witryny. Jest to przydatne dla użytkowników oraz sprawia, że strona internetowa staje się łatwiejsza do znalezienia przez wyszukiwarki i do archiwizacji[[6]](#footnote-6).
* Należy upewnić się, że zawartość nie jest dodawana i usuwana między sesjami akwizycji, ponieważ ta dodana zawartość nie zostanie przechwycona i zachowana.
* Aby zostać przechwyconą przez mechanizmy archiwizacji w Internecie, informacja musi być "osiągalna maszynowo", co oznacza, że ​​może zostać znaleziona przez robota sieciowego. Informacje, które do dostępu do siebie wymagają zalogowania się, wykorzystania pola wyboru, listy wyboru lub pola wyszukiwania, nie są osiągalne dla maszyny, więc nie mogą zostać przechwycone przez robota sieciowego. Tam, gdzie to możliwe, należy zamieścić rozwiązania alternatywne, które można pobrać bezpośrednio, takie jak lista A-Z lub mapa witryny.
* Trzeba pamiętać, że treści publikowane na witrynach należących do stron zewnętrznych, takich jak Flickr lub YouTube, mogą również wymagać zarządzania i ochrony za pośrednictwem własnej strony internetowej lub innych systemów elektronicznych. Dzieje się tak dlatego, że te strony internetowe są poza kontrolą organizacji, do której należy strona WWW i nie zawsze można je skutecznie i całkowicie zebrać. Warto przeanalizować na przykład zarchiwizowane wersje strony internetowej premiera Wielkiej Brytanii, numer 10, które wykorzystują witryny Flickr, YouTube i Twitter: [http://webarchive.nationalarchives.gov.uk/20110131164346/http://www.number10.gov.uk/](http://webarchive.nationalarchives.gov.uk/20110131164346/http%3A//www.number10.gov.uk/).
* Nie zawsze jest jasne, gdzie znajdują się granice jednej witryny; linki zewnętrzne mogą zostać przechwycone dla jednego "przeskoku" ze strony internetowej lub użytkownik strony zarchiwizowanej może połączyć się z zewnętrzną witryną na żywo.
	+ 1. Więcej informacji o tym, jak różne aspekty stron internetowych i formatów informacji online mogą wpływać na archiwizowanie stron, można znaleźć na stronie Głównego Urzędu Informacji: http://coi.gov.uk/guidance.php?page=244 (są one ukierunkowane na centralne strony rządowe, ale mogą być przydatne przy rozważaniu opcji dla wszystkich stron internetowych).
1. Dostęp i ochrona

4.1 Jak długo i dla kogo

4.1.1 Przyczyny archiwizacji witryny internetowej dają odpowiedź na to, w jaki sposób zarządzać dostępem do gromadzonych zasobów informacji. Pozwolą również ustalić, dlaczego są one przechowywane, jak często się je gromadzi, jak długo muszą być przechowywane i kto może uzyskać do niech dostęp. Jeśli materiał został zarchiwizowany ze względów prawnych, będzie przechowywany i używany zgodnie z odnośnymi wymaganiami. Natomiast organizacje dziedzictwa, takie jak archiwa, biblioteki i muzea, tworzą archiwa internetowe na większą skalę i starają się je trwale zachować.

* + 1. Zarchiwizowane strony internetowe są zwykle udostępniane w określonej formie określonym odbiorcom - albo wyłącznie w danej organizacji, albo szerszej grupie odbiorców w celach badawczych. Dostęp zależy od powodów, dla których strony zostały zarchiwizowane i od tego, czy dostęp ma być otwarty, ograniczony czy też zamknięty. Strony internetowe zbierane do celów prawnych będą zazwyczaj dostępne i wykorzystywane wyłącznie w organizacjach, które są za to odpowiedzialne.
		2. Organizacje zajmujące się dziedzictwem zazwyczaj zapewniają również dostęp dla naukowców na miejscu i online, chyba że dostęp do treści jest ograniczony. Prawa autorskie lub licencje mogą uniemożliwić ogólny dostęp do materiałów na ogólnodostępnej stronie internetowej. Archiwizacja materiałów autorskich stron trzecich może być dozwolona na mocy ustawodawstwa, chociaż zarchiwizowane materiały są udostępniane tylko w określonych lokalizacjach (np. w czytelni biblioteki). W wielu krajach poprzez modyfikację przepisów o egzemplarzu obowiązkowym umożliwiono powszechne archiwizowanie webu. Zazwyczaj dostęp do tych zarchiwizowanych materiałów jest ograniczony do czytelni krajowych instytucji zbiorowego zarządzania.
		3. Szerszy dostęp online jest możliwy tam, gdzie pozwalają na to uprawnienia związane z daną kolekcją. Brytyjskie archiwum rządowe The National Archives jest jedną z kolekcji, dostępną bezpłatnie online: www.nationalarchives.gov.uk/webarchive/. Stanowi to korzystne rozwiązanie, ponieważ The National Archives zachęca menedżerów witryn rządowych do przekierowywania użytkowników do tego archiwum internetowego kiedy rządowe strony internetowe nie są już dostępne „na żywo”: www.nationalarchives.gov.uk/information-management/policies/web-continuity.htm Podejście to wspiera dobre praktyki w rządowym zarządzaniu informacjami i pozwala każdemu użytkownikowi w Internecie na dostęp do zarchiwizowanych treści.
		4. The National Archives gromadzi strony internetowe rządu Wielkiej Brytanii, dokumentujące zmiany w interakcjach z rządem oraz sposób komunikacji rządu z opinią publiczną. Istnieje szereg innych organizacji archiwizujących strony internetowe w Wielkiej Brytanii i na całym świecie. Szczegóły ich dotyczące są dostępne pod linkiem: [www.nationalarchives.gov.uk/webarchive/other-collections.htm](http://www.nationalarchives.gov.uk/webarchive/other-collections.htm).

Źródło: <http://www.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf>

Tłumaczenie: Marzena Marcinek

Artykuł został przetłumaczony w ramach projektu: „Upowszechnianie wiedzy o archiwizacji Webu i metodach korzystania z historycznych zasobów WWW w instytucjach publicznych i sektorze NGO” – zadanie finansowane w ramach umowy 868/P-DUN/2018 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę.

1. Wikipedia, *Web archiving* http://en.wikipedia.org/wiki/Web\_archiving [↑](#footnote-ref-1)
2. Wikipedia, *Web archiving*, definicja http://en.wikipedia.org/wiki/Web\_archiving [↑](#footnote-ref-2)
3. Te definicje zostały zapożyczone z Web Archiving, Julien Masanès (ed.), (Springer, 1998). Zob. rozdział 1. [↑](#footnote-ref-3)
4. *Internet Archive*, About webpage www.archive.org/about/about.php [↑](#footnote-ref-4)
5. Informacja o standardzie W3C: www.w3.org/ ; przegląd informacji i tutoriale: www.w3schools.com/. [↑](#footnote-ref-5)
6. XML to skrót *eXtensible Markup Language*. Więcej informacji: www.w3schools.com/xml/default.asp. [↑](#footnote-ref-6)