**Tworzenie stron internetowych przystosowanych do archiwizacji**

Biblioteka Kongresu zaleca następujące dobre praktyki, o których należy pamiętać podczas projektowania stron internetowych, aby zapewnić ich pomyślną archiwizację przez dowolną instytucję archiwizującą. Chociaż przestrzeganie tych zaleceń nie gwarantuje wysokiej jakości przechwytywania archiwów i późniejszej bezbłędnej archiwizacji strony internetowej, to jednak nie przestrzeganie ich spowoduje dodatkowe wyzwania związane z archiwizacją i przechowywaniem.

**Przestrzegaj standardów sieci i wytycznych dotyczących dostępności**

Kierowanie się wytycznymi standardów sieciowych ([https://www.w3.org/standards/)](https://www.w3.org/standards/)%20) i standardów dostępności (<https://www.w3.org/WAI/standards-guidelines/>) ułatwia lepszą archiwizację i odtworzenia strony internetowej. Ponieważ roboty sieciowe, w tym archiwalny robot-przeszukiwacz Heritrix (<https://en.wikipedia.org/wiki/Heritrix>), uzyskują dostęp do stron internetowych w podobny sposób do przeglądarki tekstowej, dostępne strony internetowe są bardziej przyjazne dla robotów sieciowych. Przestrzeganie standardów sieciowych minimalizuje łączne, specyficzne cechy, które Wayback Machine (<https://en.wikipedia.org/w/index.php?title=Wayback_Machine&oldid=475037483>) musi uwzględniać podczas prezentowania zarchiwizowanych stron internetowych ([https://en.wikipedia.org/w/index.php?title=W3C\_Markup\_Validation\_Service&oldid=449780385#Browser\_accommodation](https://en.wikipedia.org/w/index.php?title=W3C_Markup_Validation_Service&oldid=449780385%23Browser_accommodation)). Agencje rządowe mogą szczególnie chcieć zapoznać się z wytycznymi Sekcji 508.gov GSA (<https://section508.gov>).

**Bądź ostrożny w przypadku wyjątków w pliku robots.txt**

Pewne typy instrukcji wprowadzone do pliku robots.txt (<https://en.wikipedia.org/wiki/Robots_exclusion_standard>) mogą być od razu dobre dla robotów-wyszukiwarek, ale mogą też uniemożliwiać archiwalnym robotom indeksującym przechwytywanie treści, która ma kluczowe znaczenie dla wiernego odtworzenia strony internetowej. Na przykład wydanie robotom indeksującym polecenia, aby pomijały katalogi CSS i JavaScript strony internetowej, nie umniejszyłoby znacząco jakości indeksu wyszukiwarki, ale spowodowałoby dużą różnicę w jakości archiwalnego przechwytywania.

**Użyj mapy strony, klarownych linków i ciągłej nawigacji**

Robot-przeszukiwacz może przechwytywać wyłącznie te strony internetowe, których istnienia jest świadomy. Odkrywa strony internetowe, przemierzając linki, co oznacza, że ostatecznie może przechwytywać tylko strony, do których może uzyskać dostęp za pomocą linków. Następstwem tego jest fakt, że użytkownik przeglądający zarchiwizowaną stronę internetową może nawigować podążając tylko za odnośnikami, ponieważ funkcje leżące po stronie serwera, takie jak wyszukiwanie, nie działają w archiwum. Zaleca się unikanie polegania na Flash’u, rozwiązaniach typu JavaScript lub innych technikach, które mają tendencję do utajniania linków jako jedynego sposobu nawigacji do dowolnej strony i rozważenie utworzenia obszernej mapy strony (<https://en.wikipedia.org/w/index.php?title=Site_map>), aby upewnić się, że robot niczego nie przeoczy.

**Utrzymuj stabilne identyfikatory URI i przekierowuj, jeśli to konieczne**

Stabilność identyfikatora URI stron internetowych umożliwia przeglądanie przechwyconych stron internetowych, począwszy od roku 1997, ([http://webarchive.loc.gov/all/\*/www.loc.gov/](http://webarchive.loc.gov/all/*/www.loc.gov/)) w archiwum internetowym Biblioteki Kongresu. Oznacza to również, że wszelkie zakładki lub linki opublikowane i rozpowszechniane nadal działają bez zmian. Link rot (<https://en.wikipedia.org/wiki/Link_rot>), mimo przeciwieństw, jest w sieci ogólnie powszechny (<https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/>).

Kiedy identyfikator URI zmienia się i przekierowuje do nowej lokalizacji zasobu nie zmniejsza się prawdopodobieństwo, że nowy URI zostanie zarchiwizowany. Powyższe działania gwarantują, że dostęp do archiwów strony internetowej sprzed zmiany URI zostanie odłączony od stron po zmianie URI. Wrażliwość narzędzi do archiwizacji stron internetowych na stabilność URI oznacza również, że identyfikatory URI zawierające ID zewnętrznych sesji (<https://en.wikipedia.org/wiki/Session_ID>) mogą być podobnie oddzielone od wcześniejszych procesów przechwytywania tego samego zasobu.

**Rozważ korzystanie z licencji Creative Commons**

Biblioteka Kongresu musi prosić większość właścicieli stron internetowych o zgodę na ponowne wyświetlenie ich stron internetowych poza siecią biblioteczną, a nawet o zaindeksowanie ich stron. Biblioteka Kongresu jest jedną z wielu (<http://netpreserve.org/web-archiving/legal-issues/>) instytucji zewnętrznych, która archiwizując strony internetowe musi pozyskiwać uprawnienia. Opublikowanie strony internetowej na licencji Creative Commons (<https://creativecommons.org/>) oznacza zgodę na przeszukiwanie i archiwizację.

**Używaj trwałych formatów danych**

Chociaż strona internetowa jest przedstawiana jako ujednolicony produkt, składa się z wielu różnych plików i typów plików. Zobowiązanie do zachowania tego produktu oznacza zobowiązanie do zarządzania potencjalnie odrębnymi zagrożeniami związanymi z przechowywaniem wszystkich typów plików będących częściami składowymi. Kiedy wybierasz typy kodu i formaty plików używanych podczas tworzenia strony internetowej, otwarte standardy i otwarte formaty plików są zazwyczaj najlepszymi sposobami utrwalenia. Wyjątek stanowi sytuacja, gdy otwarty format jest słabo udokumentowany lub pozwala na rozszerzenia specyficzne dla danego dostawcy – mogą one być gorsze niż dobrze udokumentowane zastrzeżone formaty, które są szeroko i w jednolity sposób wdrażane. Strona internetowa poświęcona Zrównoważonemu Rozwojowi Formatów Cyfrowych (<https://www.loc.gov/preservation/digital/formats/index.html>) przedstawia szereg kryteriów ([https://www.loc.gov/search/500/#factors](https://www.loc.gov/search/500/%23factors)), które pozwalają uzyskać prawdziwie „zrównoważony” format poza pozorną „otwartością”.

**Osadź metadane, zwłaszcza kodowanie znaków**

Ponieważ serwery internetowe nie przekazują w niezawodny sposób kodowania znaków (<https://www.w3.org/blog/2008/03/html-charset/>), ważne jest, aby strony internetowe robiły to dobrze. Użyj metatag’u HTML (<https://www.w3.org/TR/html401/struct/global.html#adef-http-equiv>) lub XML do wskazania ([https://www.w3.org/TR/REC-xml/#NT-EncodingDecl](https://www.w3.org/TR/REC-xml/%23NT-EncodingDecl)), jakie kodowanie powinno być używane do prezentowania strony. Dodatkowe metadane są przydatne dla organizacji, które tworzą zbiory archiwów internetowych, takie jak te przechowywane przez Bibliotekę Kongresu, które czerpią z dostarczanych przez stronę metadanych czerpią informacje dla punktów dostępu i rekordów opisowych.

**Korzystaj z dostawców platform przyjaznych archiwizacji i z systemów zarządzania treścią**

Podczas gdy dostawcy platform, takich jak media społecznościowe czy firmy publikujące w Internecie mają motywację, aby umożliwić komercyjnym podmiotom indeksującym dostęp do przynajmniej niektórych udostępnianych treści, nie zawsze są one tak przychylne dla archiwalnych robotów indeksujących. Jeśli archiwizacja strony jest ważna, sprawdź plik robots.txt firmy lub zapytaj o ich politykę udostępniania, zanim przejdziesz na ich platformę. Ponadto, nawet jeśli firma nie blokuje natychmiast archiwalnych robotów indeksujących, stosowane szablony stron lub systemy zarządzania treścią, mogą nie archiwizować się dobrze. Zobacz, jak inne strony internetowe zbudowane na tej samej platformie odtwarzają się w archiwach internetowych, na przykład w Archiwach Internetowych Biblioteki Kongresu (<http://webarchive.loc.gov/>), a jeśli korzystasz z systemu zarządzania treścią o otwartym kodzie źródłowym, zapoznaj się z konfiguracją dowolnego pakietu plików robots.txt.

**Dodatkowe źródła**

Możesz być także zainteresowany dodatkowymi źródłami:

1. Biblioteka Kongresu - Zalecane formaty oświadczeń dla stron internetowych (<https://www.loc.gov/preservation/resources/rfs/index.html>)
2. Wytyczne archiwizacji Uniwersytetu Stanforda (<http://library.stanford.edu/projects/web-archiving/archivability>)
3. Wytyczne dla archiwizowanych stron Uniwersytetu Columbia (<https://library.columbia.edu/bts/web_resources_collection/guidelines_for_preservable_websites.html>)
4. Wytyczne Uniwersytetu Princeton dotyczące projektowania witryn przyjaznych dla środowiska (<https://rbsc.princeton.edu/policies/guidelines-designing-preservation-friendly-websites>)
5. Archiwum Ready, bezpłatne narzędzie do oceny narzędzi do archiwizacji stron WWW: wytyczne dotyczące projektowania przyjaznych dla archiwizacji stron (<http://www.archiveready.com/>)

Tłumaczenie:

Joanna Pawliczak

Źródło:

<https://www.loc.gov/programs/web-archiving/for-site-owners/creating-preservable-websites/>

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.