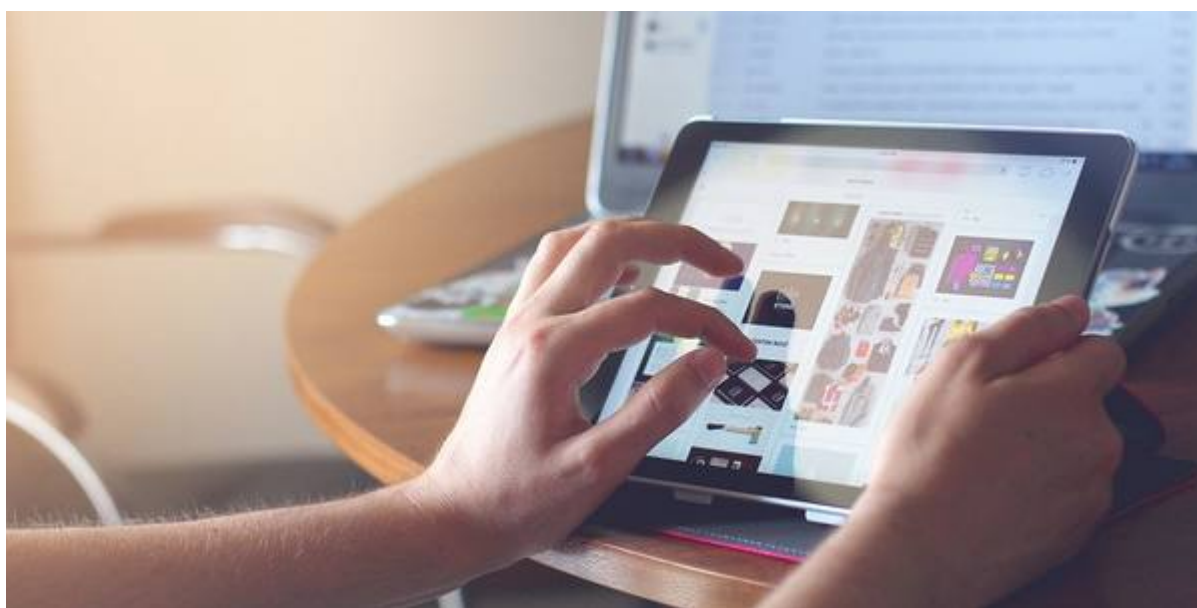


Czy archiwizacja zasobów sieci to realny problem?

14 stycznia 2018

Strony internetowe istnieją przeciętnie od 40 do 100 dni. W wielu krajach europejskich ich archiwizowaniem zajmują się biblioteki narodowe. W Polsce brakuje takiej inicjatywy, choć problem ten dostrzegają m.in. naukowcy i prawnicy – powiedział PAP ekspert z UW.



„Nie jesteśmy w stanie zarchiwizować wszystkich zasobów internetu, jest ich po prostu za dużo. Jednak podejmowane są próby zapisania dla przyszłych pokoleń części stron internetowych. Niektóre instytucje dbają też o archiwizację zawartości swoich skrzynek mailowych i oficjalnych kont w mediach społecznościowych” – powiedział PAP ekspert z Laboratorium Cyfrowego Humanistyki UW Marcin Wilkowski.

Każdej minuty internauci dzielą się Snapchacie ponad pół milionem zdjęć, umieszczają blisko pół miliona wpisów na Twitterze, prawie 50 tys. zdjęć na Instagramie; sami tylko Amerykanie używają ponad 2,6 mln GB danych. Aż 90 proc. dostępnych danych powstało w ciągu ostatnich dwóch lat. Dziennie pojawia się ich 2,5 kwintyliona bajtów – wynika z

raportu Data Never Sleeps 5.0 przygotowanego przez firmę DOM0.

W wielu krajach europejskich, m.in. w Niemczech, Austrii, Czechach, Finlandii, Wielkiej Brytanii, archiwizowaniem zawartości stron internetowych zajmują się biblioteki narodowe. Nie rejestrują jednak wszystkich stron, a jedynie te, które opublikowane są w domenie danego kraju (czyli w Niemczech z końcówką „de”, a w Wielkiej Brytanii – „uk” i „co.uk”). Inne kraje oprócz stron w domenie narodowej archiwizują również ważne publikacje internetowe, które ich dotyczą. Tak robi na przykład Portugalia w swoim archiwum webowym, powstałym już 10 lat temu – wskazuje Wilkowski.

„W Polsce brakuje podobnej inicjatywy” – zaznaczył ekspert. „Nigdzie nie jest i nigdy nie będzie dostępny pełen zasób archiwalnych stron internetowych. Nie jest zatem prawdą, że w internecie nic nie ginie” – uważa Wilkowski. Przypomniał, że przeciętny czas istnienia strony internetowej wynosi od 40 do 100 dni, a dodatkowo niektóre rozwiązania w projektowaniu stron internetowych, takie jak JavaScript, utrudniają ich archiwizację.

Jak poinformowała PAP Malwina Rozwadowska z Narodowego Archiwum Cyfrowego (NAC) w latach 2009-2010 NAC przeprowadziło przedsięwzięcie polegające na archiwizacji stron internetowych należących do domeny „gov.pl”. „Był to jednorazowy projekt. Obecnie trwają rozmowy z Ministerstwem Cyfryzacji nad kontynuacją archiwizacji internetu, jednak nie jesteśmy na tym etapie w stanie podać konkretnych informacji co do daty oraz formy udostępnienia danych dla szerszej publiczności” – podkreśliła Rozwadowska.

Ponad 279 mld stron WWW archiwizowanych od 1996 roku dostępnych jest dziś dzięki amerykańskiej fundacji Internet Archive, zajmującej się także digitalizacją, udostępnianiem zbiorów multimedialnych czy starych gier komputerowych.

„Założyciel Internet Archive, Brewster Kahle, porównał jej

działalność o Biblioteki Aleksandryjskiej, której celem była próba zebrania wszelkich możliwych tekstów pisanych z całego świata. Już w połowie lat dziewięćdziesiątych strony internetowe Kahle uznawał za część dziedzictwa cyfrowego” – opowiada Wilkowski. W 2002 r. Internet Archive podpisało ze współczesną Biblioteką Aleksandryjską porozumienie, na mocy którego ta druga zobowiązała się do stworzenia kopii zapasowej archiwizowanych zbiorów internetu.

W dyskusję o archiwizacji żywo włączają się również prawnicy. Podczas rozpraw sądowych coraz częściej w materiale dowodowym pojawiają się bowiem odesłania do stron internetowych, również archiwalnych, które już nie istnieją. Problemem, zwłaszcza w tym drugim przypadku, jest uznanie w procesach sądowych ich wiarygodności. Podobnie naukowcy mają problem ze znikającymi stronami, do których odwołują się w przypisach autorzy prac naukowych.

„Podczas gdy zarchiwizowanie strony internetowej jest zadaniem stosunkowo prostym, zdecydowanie trudniej jest z zawartością mediów społecznościowych, które stają się dziś przestrzenią oficjalnej komunikacji instytucji publicznych” – uważa Wilkowski. Dlatego w niektórych państwach archiwizowane są ministerialne tweety. Tak dzieje się na przykład w Wielkiej Brytanii.

Dużym wyzwaniem dla archiwistyki jest też to, że Internet jest dziś w dużej mierze spersonalizowany i treść wielu stron internetowych dostosowywana jest dynamicznie do wcześniejszych wyborów użytkownika. Oznacza to, że tym samym adresem URL różne osoby mogą otrzymać różne treści; co w takim przypadku jest oryginałem, który należałoby zabezpieczyć? – zastanawia się Wilkowski.

W ocenie Wilkowskiego próby archiwizacji wpisów zamieszczanych w mediach społecznościowych są dla osób postronnych mocno ograniczone przez regulaminy platform i limity danych, które można pozyskać przez specjalne interfejsy programistyczne. Do

tego dochodzi skala zapisywanych danych.

Biblioteka Kongresu w USA zaczęła gromadzić wszystkie wiadomości z Twittera w kwietniu 2010 roku i do tej pory zarchiwizowała kilka miliardów tweetów. W grudniu 2017 r. ogłosiła, że już nie będzie gromadziła wszystkich wpisów publikowanych w tym serwisie społecznościowym. Począwszy od 1 stycznia 2018 r. instytucja ta selekcjonuje wiadomości, które zachowa na przyszłość. Będą to ważne wydarzenia i tendencje społeczne.

Internet to również skrzynki mailowe, w tym takie, których używają oficjalne instytucje państwowe czy głowy państw. Na przykład w USA archiwizowane są skrzynki mailowe prezydentów. „Niestety, polskie instytucje publiczne nie ujawniają zasad i metod archiwizacji swoich zasobów internetowych” – uważa Wilkowski.

Malwina Rozwadowska z NAC poinformowała PAP, że instytucja ta nie zajmuje się archiwizacją skrzynek mailowych polskich instytucji publicznych.

Kancelaria Prezydenta RP – na pytanie PAP, o to, czy oficjalna prezydencka skrzynka mailowa i jego ministrów są archiwizowane – odpowiedziała, że brak jest uzasadnienia do takiego działania. „Archiwizacja danej wiadomości następuje więc o tyle, o ile stanowi ona wszczęcie sprawy np. skargowo – wnioskowej albo część sprawy, na skutek których to okoliczności zostaje włączona do akt sprawy. Archiwizowanie wiadomości nie stanowiących części sprawy byłoby więc bezprzedmiotowe” – informuje kancelaria Prezydenta RP. „Należy także mieć na uwadze, że każdy z użytkowników poczty elektronicznej co do zasady samodzielnie zarządza zawartością swojej skrzynki pocztowej” – dodała kancelaria.

Do publikacji materiału PAP nie otrzymała odpowiedzi od KPRM na pytanie dot. ewentualnej archiwizacji skrzynek rządowych – premiera, ministrów, czy tych należących do wojewodów.

Zdaniem Marcina Wilkowskiego najbardziej racjonalną postacią archiwizacji internetu jest zabezpieczanie zawartości witryn powstających w domenie narodowej oraz zasobów mediów społecznościowych dokumentujących ważne wydarzenia – np. w przypadku Polski wpisów dot. np. Światowych Dni Młodzieży czy wyborów. Taki zasób musi być gromadzony już w czasie trwania wydarzeń.

Pytany o przyszłość badania zasobów internetu – stwierdził, że od lat rozwijane są ich metody. „Nie da się jednak wykonywać badań historycznych dotyczących internetu bez programowania i narzędzi cyfrowych. Te kompetencje staramy się rozwijać w powołanym w 2015 roku na UW Laboratorium Cyfrowym Humanistyki UW” – powiedział.

Autorstwo: Szymon Zdziebłowski

Zdjęcie: [fancycravel](#) (CC0)

Źródło: [NaukawPolsce.PAP.pl](#)