
Cytowania zasobów sieciowych w polskich czasopismach z zakresu bibliotekoznawstwa i informatologii: analiza aktualności adresów URL

Marcin Roszkowski

*Katedra Informatologii, Wydział Dziennikarstwa, Informacji i Bibliologii,
Uniwersytet Warszawski*

Bartłomiej Włodarczyk

*Katedra Bibliografii i Dokumentacji, Wydział Dziennikarstwa, Informacji i Bibliologii,
Uniwersytet Warszawski*

Abstrakt

Cel/teza: Celem artykułu jest charakterystyka stanu badań nad problemem dostępności odwołań do zasobów sieciowych obecnych w aparacie pomocniczym tekstów naukowych oraz przedstawienie wyników badań własnych związanych z tym zagadnieniem w odniesieniu do artykułów naukowych z czterech polskich czasopism z zakresu bibliotekoznawstwa i informatologii.

Koncepcja/metody badań: Wykorzystując krytyczną analizę piśmiennictwa naukowego zidentyfikowano grupy zagadnień szczegółowych będących przedmiotem zainteresowania badaczy. Wskazano również na metodologiczne aspekty badania tego zjawiska oraz nakreślono jego obraz w odniesieniu do poszczególnych dyscyplin naukowych. Przedmiotem badań empirycznych były artykuły naukowe opublikowane w czterech polskich czasopismach z zakresu bibliotekoznawstwa i informatologii w latach 2009–2015: *Bibliotheca Nostra*, *Śląski Kwartalnik Naukowy*, *Biuletyn EBIB*, *Przegląd Biblioteczny*, *Zagadnienia Informacji Naukowej – Studia Informacyjne*. Dostępność adresów URL w cytowaniach zbadano metodą automatyczną za pośrednictwem kodów odpowiedzi serwerów HTTP.

Wyniki i wnioski: W czasopismach uwzględnionych w badaniu ok. 37% wszystkich pozycji w bibliografiach załącznikowych i przypisach zawierało łącza do zasobów internetowych. Wyniki badań nad dostępnością zasobów sieciowych cytowanych w korpusie analizowanych publikacji wskazują, że ponad 70% odwołań jest nadal aktywnych w momencie przeprowadzenia badania. Wartość tej zmiennej jest zróżnicowana dla poszczególnych tytułów. Wskaźnik połowicznej dezaktywacji łączy dla czterech zbadanych tytułów wynosi ok. siedmiu lat.

Oryginalność/wartość poznawcza: Przedstawiony stan badań związanych z aktualnością odwołań do zasobów sieciowych w aparacie pomocniczym publikacji naukowych może być punktem odniesienia dla badań szczegółowych związanych z tym zagadnieniem i być źródłem rozwiązań metodologicznych. Problem ten nie był według wiedzy autorów eksplorowany w odniesieniu do polskich czasopism z zakresu bibliotekoznawstwa i informatologii. Tym samym przedstawione wyniki są pierwszą próbą charakterystyki skali tego zjawiska.

Słowa kluczowe

Analiza cytowań. Polskie czasopisma naukowe. Czasopisma z zakresu bibliotekoznawstwa i informatologii. Wskaźnik połowicznej dezaktywacji łączy. Cytowania zasobów sieciowych.

Otrzymano: 3 lipca 2016. Zrecenzowano: 12 lipca 2016. Zaakceptowano: 18 lipca 2016.

1. Wprowadzenie

Przypisy bibliograficzne oraz bibliografie załącznikowe stanowią istotny element pracy naukowej. Służą one dokumentacji postawionych tez i powinny umożliwić innym badaczom sprawdzenie poprawności powołań i całego wywodu naukowego. W związku ze wzrostem znaczenia zasobów sieci World Wide Web (WWW) w obiegu informacji w nauce, odwołania do zasobów sieciowych są niemal stałym elementem w bibliografiach załącznikowych publikacji naukowych. Istotnym problemem w przypadku tych materiałów pozostaje kwestia ich dostępności i stabilności. Badania nad aktywnością hiperłączy i trwałością zasobów internetowych rozpoczęły się wkrótce po powstaniu WWW, w połowie lat 90. XX w. Wiązały się one ze świadomością, że WWW nie tylko rozrasta się, ale także część jego zasobów znika, przestaje być dostępna (tzw. zjawisko *link rot*) pod podanym adresem Uniform Resource Locator (URL). W części zaś przypadków zmiana ulega treść stron internetowych (ang. *content drift*). Badania prowadzone na przełomie XX i XXI w. wskazywały na dużą skalę zjawiska zanikania zasobów WWW. Wallace Koehler pokazał na losowej próbie 361 stron WWW, zebranej w grudniu 1996 r., że problem zwiększa się wraz z upływem czasu. W styczniu 2001 r. aktywne były tylko 124 strony (34.4% stanu wyjściowego), a w maju 2003 r. – 122 (33.8%). Badacz zaobserwował zjawisko wznawiania działalności stron, które przez pewien okres były nieaktywne (Koehler, 2004). Najnowszy przegląd wspomnianej próbki stron WWW pokazał, że pod koniec 2013 r. aktywnych było tylko 6 łączy (1.6% stanu wyjściowego), a w lutym 2015 r. na zapytanie odpowiedziały tylko dwie strony (Oguz & Koehler, 2016, 477).

Od początku tego wieku przeprowadzono liczne badania dotyczące łączy do zasobów internetowych zamieszczanych w przypisach bibliograficznych i bibliografiach załącznikowych publikacji naukowych. W polskiej literaturze z zakresu bibliotekoznawstwa i informatologii (BIN) brakuje badań poświęconych tej problematyce. Artykuł ma na celu uzupełnienie tej luki w polskim piśmiennictwie naukowym. Postawiono sobie w nim dwa cele, zaprezentowanie stanu badań oraz przedstawienie wyników badania cytowania zasobów sieciowych w wybranych czasopismach ze wspomnianej dziedziny.

Artykuł został podzielony na dwie części. W części pierwszej przedstawiono stan badań nad cytowaniem zasobów sieciowych oraz propozycje rozwiązań problemów dotyczących trwałości cytowanych zasobów internetowych. Skoncentrowano się na badaniach dotyczących literatury z zakresu BIN. Następnie w drugiej części artykułu zaprezentowano metodologię badania oraz uzyskane rezultaty. W zakończeniu wskazano na propozycje rozwiązań, które warto poddać pod dyskusję w kontekście wdrożeń w polskim środowisku wydawców i autorów prac naukowych.

2. Stan badań

W ramach badań nad łącami internetowymi będącymi elementami bibliografii załącznikowych można wyróżnić dwie podstawowe grupy. Pierwsza z nich zawiera badania ilościowe przypisów i bibliografii załącznikowych zawierających łącza do zasobów internetowych, druga zaś koncentruje się na propozycjach rozwiązań zapewniających większą trwałość cytowanych zasobów.

Jednym z celów badań sytuujących się w ramach grupy pierwszej jest wskazanie liczby aktywnych i nieaktywnych łączy oraz przyczyn tego stanu. Badania dotyczą linków zamieszczonych w jednym czasopiśmie, kilku czasopismach z jednej dziedziny lub porównania charakterystyki łączy w czasopismach z różnych dziedzin. W ramach tej grupy badań prowadzone są również prace dotyczące zmiany treści zasobów w czasie. Przywoływane strony WWW mogą być dostępne, jednak ich zawartość może różnić się od prezentowanej w momencie sporządzania przypisu. Wallace Koehler i Fatih Oguz, którzy odróżnili badania dotyczące aktywności łączy od tych skupiających się na zmianie treści zasobów, podkreślili, że dużo więcej jest prac koncentrujących się na pierwszym zagadnieniu (Oguz & Koehler, 2016, 477).

W pierwszej grupie badań można wyróżnić następujące podstawowe zagadnienia badawcze:

- (1) Liczba cytowań zasobów internetowych.
- (2) Stosunek liczby łączy nieaktywnych do aktywnych (w tym zmiana treści przywoływanych zasobów).
- (3) Rodzaje odpowiedzi serwera.
- (4) Wpływ długości adresu URL na czas aktywności łączy.
- (5) Wpływ rodzaju domeny na czas aktywności łączy.

Wyniki badań wskazują na wzrost liczby powołań na zasoby internetowe w czasie. Dotyczy to zarówno ogólnej liczby cytowań materiałów sieciowych, jak i średniej liczby cytowań tych zasobów na artykuł. Nie jest to jednak wzrost liniowy. Niektóre badania wskazują na zahamowanie szybkiego trendu wzrostowego zaobserwowane w ostatnich latach. Podobne wyniki uzyskano badając publikacje z różnych dziedzin nauki. W porównywaniu wyników badań istotnym czynnikiem jest czas, z którego pochodzą analizowane dane. W związku z tym obok wyników zamieszczono w nawiasach okrągłych okres, z którego pochodzą badane zasoby. Wspomniany trend jest widoczny w badaniu przeprowadzonym przez Siluo Yanga, Ruizhen Hana, Jingda Dinga oraz Yanfei Songa. Chińscy naukowcy analizowali trzy zbiory danych dotyczące cytowań z zakresu humanistyki i nauk społecznych (1998–2009), informatyki (1995–1999) oraz medycyny (1994–2006). Rozkład cytowań zasobów internetowych przypomina we wszystkich dziedzinach literę „S”. Po okresie szybkiego wzrostu następuje jego wyhamowanie. Różnice między zbiorami dotyczą m.in. liczby cytowanych materiałów internetowych na artykuł. Jedną z przyczyn zaobserwowanej odrębności może być pochodzenie materiału badawczego z różnych dziedzin wiedzy. Zdaniem badaczy, zaobserwowane zahamowanie wzrostu liczby analizowanych powołań wynika z ograniczeń dotyczących liczby cytowanych pozycji bibliograficznych w jednym artykule i liczby artykułów w pojedynczym czasopiśmie (Yang et al., 2012, 782–783). Dion Hoe-Lian Goh oraz Peng Kin Ng również wskazali na wzrost liczby cytowanych materiałów internetowych w trzech czasopismach z zakresu BIN (1997–2003). W 1997 r. artykuł zawierał średnio 0.89 cytowania zasobu internetowego, w 2002 r. – 6.42, jednak w 2003 r. było tych powołań średnio tylko 3.93. Badacze nie byli w stanie przedstawić pewnej przyczyny spadku liczby powołań na zasoby internetowe w ostatnim analizowanym roku (Goh & Ng, 2007, 18–19).

Zbieżne wyniki uzyskiwano w odniesieniu do stosunku między liczbą łączy aktywnych a liczbą łączy nieaktywnych w przypisach i bibliografiach załącznikowych. Liczba aktywnych łączy była odwrotnie proporcjonalna do ich wieku. Zależność taką zaobserwowano w przypadku czasopism z wielu dziedzin nauki, m.in. z prawa (Zittrain et al., 2014, 185),

medycyny (Habibzadeh, 2013, 457), BIN (Kumar et al., 2015, 61; Wu, 2008, 487) oraz informatyki biomedycznej (Carnevale & Aronsky, 2007, 271). Ten generalny wzrostowy trend nie oznacza identyczności wyników wszystkich badań. Różnią się one między sobą nie tylko w poszczególnych dziedzinach, ale także w ramach jednej dyscypliny naukowej. Również ranga czasopisma może mieć wpływ na procent aktywnych i dostępnych URL w cytatach. P. Habibzadeh pokazał, że w przypadku czasopism medycznych (2006–2013), trwałość łączy jest większa w międzynarodowych, prestiżowych publikacjach z tej dziedziny. Zdaniem badacza niższa trwałość łączy w czasopismach regionalnych może wynikać z ograniczonego dostępu naukowców w krajach rozwijających się do naukowych baz danych i innych zasobów (Habibzadeh, 2013, 458). Poszczególne badania różnią się między sobą przyjętą metodologią. Niektóre z nich uwzględniają zmianę treści zasobów poprzez zaliczanie linków prowadzących do zmienionych treści do grupy łączy nieaktywnych (zob. np. Wu, 2008), inne zaś ignorują tę kwestię (zob. np. Goh & Ng, 2007). Na przykład Zhiqiang Wu, badający cytowania materiałów internetowych w dwóch chińskich czasopismach z zakresu informatyki i BIN (1999–2003), podkreślił, że krzywa wskazująca liczbę łączy kierujących do zasobów o niezmienionej treści jest zawsze poniżej krzywej pokazującej liczbę aktywnych linków. W 2003 r. stosunek ten wynosił 321 do 449 URL (Wu, 2008, 486–487). W badaniach Jonathana Zittraina, Kendry Alberta, Lawrence’a Lessiga z 353 wybranych aktywnych łączy z korpusu opinii Sądu Najwyższego Stanu Zjednoczonych (1996–2011) tylko 76% prowadziło do zasobów niezmienionych treściowo (Zittrain et al., 2014, 182). Zjawisko rozbieżności między liczbą aktywnych łączy a liczbą łączy prowadzących do cytowanych przez autorów zasobów było analizowane również odnośnie czasopism z zakresu BIN (zob. Mardani & Sangari, 2013, 35–36).

Badacze starali się również odnajdywać zasoby niedostępne pod adresem umieszczonym w przypisie bądź bibliografii załącznikowej. W tym celu wykorzystywali najczęściej wyszukiwarki internetowe, serwis Internet Archive (Wayback Machine) oraz samodzielnie szukali cytowanych materiałów w serwisach internetowych. Mary F. Casserly oraz James E. Bird analizujący próbkę 500 przypisów z czasopism z zakresu BIN (1999–2000), uzyskali dzięki Internet Archive dostęp do 47% niedostępnych już, a cytowanych wcześniej zasobów (Casserly & Bird, 2003, 310). Wu udało się odnaleźć aż 74.4% brakujących dokumentów. Z tego 62.8% uzyskano wykorzystując wyszukiwarki, 24.4% serwis Internet Archive, a 12.8% przeszukując oryginalne strony WWW zawierające cytowane materiały (Wu, 2008, 494–495). D. Vinay Kumar, B. T. Sampath Kumar oraz D. R. Parameshwarappa uwzględnili w swoim badaniu zależność między różnymi analizowanymi czynnikami, np. długością łączy a możliwością odszukania zasobów w Internet Archive (Kumar et al., 2015). Zaprezentowane rezultaty wskazują, że większość cytowanych zasobów nadal istnieje w Internecie. Zmianie uległa tylko ich lokalizacja w sieci WWW. Pokazuje to zasadność utrzymywania stałych identyfikatorów odsyłających do materiałów sieciowych. Kwestia ta zostanie opisana dokładniej w dalszej części artykułu.

Porównanie wyników uzyskanych przez różnych badaczy umożliwia współczynnik wskazujący na czas połowicznej dezaktywacji łączy (ang. *half-life*). Pokazuje on czas po jakim połowa cytowanych linków nie będzie aktywna. Badacze wyliczali wartość tego współczynnika zarówno dla konkretnych czasopism, jak i dziedzin wiedzy. W tym drugim przypadku czas połowicznej dezaktywacji łączy był równy średniej uzyskanej dla kilku czasopism z danej dziedziny. Współczynnik ten jest odmienny dla różnych dziedzin

wiedzy i czasopism (Koehler, 2004), jednak nawet w obrębie jednej dziedziny jego wartość różni się niekiedy znacznie dla konkretnych publikacji. Goh i Ng (1997–2003) oraz K. R. Prithviraj i B. T. Sampath Kumar (2001–2010) wyliczyli jego wartość dla BIN na około 5 lat (Goh & Ng, 2007, 20; Prithviraj & Kumar, 2014, 44), natomiast A. H. Mardani i M. Sangari dla irańskich czasopism z tej dziedziny (2006–2010) uzyskali wynik wynoszący 3.63 roku (Mardani & Sangari, 2013, 34–35). Jednocześnie dla internetowego czasopisma „D-Lib Magazine” (1995–2004) uzyskano wynik 10 lat (McCown, Sheffan, Nelson, & Bollen, 2005, 5), a dla czasopisma „Information Research” (1995–2008) – 14.04 lata (Moghaddam, Saberi, & Esmaeel, 2010, 71–72). Daniela V. Dimitrova i Michael Bugeja dla czasopism z zakresu dziennikarstwa i komunikacji (2000–2003) uzyskali czas połowicznej dezaktywacji na poziomie 3.17 lat (Dimitrova & Bugeja, 2007, 820), natomiast Habibzadeh dla prestiżowych międzynarodowych czasopism medycznych uzyskał wynik 4.7 lata, a dla publikacji regionalnych – 3.5 roku (Habibzadeh, 2013, 463).

Zestawiano również kody odpowiedzi Hypertext Transfer Protocol (HTTP) uzyskane w przypadku nieaktywnych łączy. Najczęstszym z nich był kod 404 wskazujący na brak możliwości odnalezienia poszukiwanego pliku. Odsetek takich odpowiedzi wahał się w przypadku badań dotyczących literatury naukowej z zakresu BIN od 36.52% (Kumar et al., 2015, 61–62) do 63.84% (Sampath Kumar & Prithvi Raj, 2012, 25). W pierwszym z wymienionych badań, dotyczącym dwóch czasopism z okresu 2008–2012, najczęściej pojawiającą się odpowiedzią był kod 500, oznaczający wewnątrz błąd serwera (39.18%). Autorzy nie odnieśli się jednak do przyczyn tej odmienności (Kumar et al., 2015, 61–62). W pozostałych badaniach kod 500 pojawiał się najczęściej na drugim miejscu, po nim zaś kod 403 oznaczający dostęp zabroniony. Poza tym wśród zwracanych oznaczeń pojawiały się kody 502, 504, 401 oraz 400 (Goh & Ng, 2007, 19; Gul, Mahajan, & Ali, 2014, 32; Kumar et al., 2015, 61–62; McCown et al., 2005, 5; Moghaddam et al., 2010, 70; Prithviraj & Kumar, 2014, 41; Sadat-Moosavi, Isfandyari-Moghaddam, & Tajeddini, 2012, 186–187; Sampath Kumar & Prithvi Raj, 2012, 25).

Diomidis Spinellis przedstawił hipotezę mówiącą, że wraz ze wzrostem głębokości adresu URL, czyli liczby katalogów w łączy, wzrasta liczba niedostępnych zasobów (Spinellis, 2003, 75). Naukowcy uwzględniają w swoich badaniach analizę związków między głębokością łączy a czasem ich aktywności. W badaniu Goha oraz Nga liczba nieaktywnych adresów URL wzrastała od 25.6% w przypadku poziomu 0 do 39.4% w przypadku poziomu 5. Goh i Ng posługując się współczynnikiem korelacji Pearsona nie stwierdzili jednak zależności między badanymi zmiennymi (Goh & Ng, 2007, 21). Istotny związek między nimi stwierdzili natomiast Prithviraj i Kumar. Ich badanie pokazało na niemożność odnalezienia od 25.44% zasobów (poziom 0) aż do 64.16% (poziom 5) (Prithviraj & Kumar, 2014, 43). Kwestią rzadko poruszaną w badaniach dotyczących literatury z zakresu BIN jest problem zależności między liczbą znaków adresu URL a trwałością łączy. Zbadali ją Kumar, Kumar oraz Parameshwarappa wskazując na większy poziom niedostępności zasobów identyfikowanych za pomocą dłuższych URL. Badacze nie przeprowadzili analizy korelacji zmiennych (Kumar et al., 2015, 63–64).

Kolejnym czynnikiem, który badano był rodzaj domeny i jego związek ze stopniem nieaktywności łączy. W badaniach dotyczących literatury z zakresu BIN stosunkowo najczęściej cytowanymi domenami były domeny *com* (*co*), *org* i *edu* (*ac*) (Goh & Ng, 2007, 22; Gul et al., 2014, 32; Moghaddam et al., 2010, 67; Prithviraj & Kumar, 2014, 42; Sampath Kumar &

Prithvi Raj, 2012, 26). W pracy Mardaniego i Sangariego, której celem była analiza przypisów z sześciu irańskich czasopism BIN, wśród często cytowanych domen występuje też domena *gov* (Mardani & Sangari, 2013, 36). Popularność domeny *com* można wiązać ze stopniem jej rozpowszechnienia w Internecie, domena *edu* jest związana z treściami naukowymi i edukacyjnymi (Goh & Ng, 2007, 22), natomiast domena *org* wydaje się być popularna ze względu na wiarygodność treści (Sampath Kumar & Prithvi Raj, 2012, 25). Wśród nieaktywnych łączy duży odsetek stanowią zazwyczaj w różnych badaniach URL z domeną *edu*. Goh i Ng stwierdzają, że taki wynik może w pierwszej chwili wydawać się niezrozumiały, ponieważ uniwersytety dysponują zazwyczaj dobrą infrastrukturą oraz wiedzą techniczną. Przyczyn takiego wyniku upatrują w dynamicznym charakterze zawartości akademickich stron WWW, w przypadku których stara zawartość jest często zastępowana nową (Goh & Ng, 2007, 22). Między czasopismami z różnych dziedzin wiedzy zachodzą różnice w liczebności cytowanych adresów z poszczególnych domen (Yang et al., 2012, 783–784).

Druga grupa badań dotyczy sposobów radzenia sobie z problemem nietrwałości łączy i zasobów internetowych cytowanych w pracach naukowych. W jednym z najwcześniejszych artykułów odnoszących się do BIN, Casserly oraz Bird zwrócili uwagę na konieczność formułowania instrukcji dla autorów, tak, aby zawierały wskazówki jak ocenić jakość i trwałość zasobów. Badacze podkreślali, że niezbędne jest uwzględnienie sposobów cytowania zasobów internetowych, które powinny zawierać jak najpełniejsze informacje na temat dokumentu, w tym datę dostępu. Redaktorzy powinni przeglądać i korygować instrukcje dla autorów, dostosowując je do aktualnych trendów (Casserly & Bird, 2003, 315–316). Zalecenia te nie straciły swojej aktualności. W związku z rozpowszechnieniem i wzrostem liczby zasobów naukowych w postaci cyfrowej, coraz więcej redakcji tworzy własne lub wykorzystuje istniejące style cytowań odnośnie dokumentów internetowych. Niektórzy wskazywali także, że lepiej jest cytować zasoby z kolekcji cyfrowych takich jak biblioteki cyfrowe czy komercyjne bazy danych, niż ze stron WWW (Goh & Ng, 2007, 23). Nie zawsze jest to jednak możliwe i wydaje się, że niepotrzebnie ogranicza możliwości autorów prac naukowych. W niektórych przypadkach rozwiązaniem może być stosowanie stałych identyfikatorów takich jak Persistent Uniform Resource Locator (PURL) lub Digital Object Identifier (DOI) (zob. np. Gul et al., 2014, 36). Przegląd 1.8 miliona tekstów opublikowanych w latach 1997–2012 w bazach: arXiv.org, Elsevier i PubMed Central pokazał, że autorzy w dużym stopniu używają Uniform Resource Identifier (URI) zamiast DOI w bibliografiach załącznikowych i przypisach. Ostatecznie po wyłączeniu zasobu Elsevier uzyskano 397 412 DOI, 505 647 identyfikatorów, które powinny być zapisane jako DOI oraz 737 847 identyfikatorów niewyrażanych w opisywanym systemie. Autorzy badania, Herbert Van de Sompel, Martin Klein i Shawn M. Jones podkreślili, że nie znają przyczyny, dla której odsetek identyfikatorów należących do drugiej kategorii jest tak wysoki. Postawili hipotezę, że jest to związane ze sposobem działania programów do zarządzania bibliografią, które pobierają lokalizujące URI zamiast identyfikującego (Van de Sompel, Klein, & Jones, 2016). Jedną z przyczyn może być również nieznanostwo systemu DOI i korzyści wynikających z jego stosowania. Należy podkreślić, że rozwiązanie to jest trudne do wprowadzenia w przypadku stron internetowych, np. blogów, ponieważ wymaga sporego wysiłku od osób tworzących i utrzymujących dany serwis (Zittrain et al., 2014, 189).

Oprócz wspomnianych wyżej sposobów, innym rozwiązaniem jest archiwizacja cytowanych zasobów internetowych. Najprostszym technicznie sposobem przywoływanym

w literaturze przedmiotu, jest ich autoarchiwizacja przez autorów publikacji naukowych oraz redakcje (Prithviraj & Kumar, 2014, 45). Problemem wydaje się w tym przypadku stabilność takiej formy zachowania treści cyfrowych. Kolejnym sposobem jest skorzystanie z największego obecnie rozwijanego serwisu do archiwizacji zasobów Internetu – Internet Archive. Pozwala on zapisać na żądanie użytkownika konkretną stronę WWW dzięki opcji „Save Page Now” (Internet Archive, b.d.). Wątpliwości może jednak budzić stabilność tego serwisu związana z prywatnym charakterem instytucji (non profit) i jej dobrowolnym wsparciem (Zittrain et al., 2014, 190). Wspomniano, że w części prac serwis ten był wykorzystywany do odszukania treści niedostępnych pod zamieszczonym w bibliografii łączem URL. Badania Ke Zhoua, Claire Grover, Richarda Tobina oraz Martina Kleina przeprowadzone na kolekcji 648 388 artykułów z publikacji wydawnictwa Elsevier, pokazało, że z 70 270 nieaktywnych łączy, archiwizacji w wybranych aktualnie funkcjonujących archiwach internetowych poddano tylko 43 745 linków (62.3%). W celu zwiększenia odsetka zarchiwizowanych zasobów, badacze zaproponowali automatyczny sposób przewidywania dezaktywacji łączy w pracach naukowych oparty na uczeniu maszynowym (Zhou et al., 2015). W sieci WWW istnieją specjalne narzędzia przeznaczone do archiwizacji materiałów internetowych cytowanych w pracach naukowych. Najpopularniejsze z nich to serwisy WebCite (WebCite, b.d.), archive.is (archive.is, b.d.) oraz Perma (Perma.cc, b.d.). Jonathan Zittrain, Kendra Albert oraz Lawrence Lessig, autorzy artykułu dotyczącego serwisu Perma, opisali przyczyny powstania narzędzia wynikające ze słabości innych dostępnych rozwiązań. Zwrócili uwagę na problemy związane z finansowaniem serwisu WebCite oraz brakiem wsparcia instytucjonalnego dla archive.is. Ponadto, oba rozwiązania bazują na jednym źródle utrzymania serwisów, co nie zapewnia wystarczającej stabilności rozwiązania (Zittrain et al., 2014, 190–191). Inne rozwiązanie zostało wdrożone w przypadku serwisu Perma, który jest utrzymywany przez sieć współpracujących ze sobą instytucji, głównie amerykańskich bibliotek prawniczych (Zittrain et al., 2014, 191). Narzędzie jest przeznaczone głównie dla instytucji amerykańskich, jednak każdy może utworzyć darmowe konto umożliwiające archiwizację 10 stron miesięcznie za darmo (Perma.cc Docs, b.d.).

3. Metodologia

Przedmiotem badań przeprowadzonych przez autorów niniejszego artykułu był zbiór artykułów opublikowanych w latach 2009–2015 w czterech czasopismach z zakresu BIN. Były to trzy czasopisma naukowe – *Bibliotheca Nostra. Śląski Kwartalnik Naukowy, Przegląd Biblioteczny, Zagadnienia Informacji Naukowej – Studia Informacyjne* i jedno fachowe – *Biuletyn EBIB*. W badaniu nie uwzględniono danych z ostatniego (4) zeszytu czasopisma *Bibliotheca Nostra. Śląski Kwartalnik Naukowy* z 2015 r. w związku z jego niedostępnością w momencie prowadzenia prac badawczych. Zbiór publikacji tworzą wyłącznie artykuły o charakterze badawczym lub przeglądowym, w których autorzy realizują przyjęte cele badawcze posługując się aparatem naukowym w postaci wykazu cytowanych publikacji. Tym samym przedmiotem badania były teksty opublikowane we wskazanych czasopismach, które zamieszczono w działach lub sekcjach wskazujących na ich badawczy charakter. Nie uwzględniono m.in. recenzji oraz sprawozdań. Łącznie przeanalizowano 670 artykułów, których pełne bibliografie załącznikowe i przypisy manualnie przeniesiono do arkusza

kalkulacyjnego. Nie uwzględniono adresów URL zamieszczonych w tekstach publikacji. Ten sposób odwołań do zasobów internetowych był stosowany w niektórych numerach *Biuletynu EBIB*. Następnie oznaczono pozycje, w których wystąpiły odwołania do zasobów sieciowych i w sposób automatyczny wyodrębniono występujące tam adresy URL. Uzyskano w ten sposób zbiór 4867 adresów URL, co po usunięciu duplikatów dało 4593 unikalnych odesłań.

Podczas przygotowywania danych do analizy natrafiono na kilka problemów. Pierwszy z nich dotyczył formatowania niektórych adresów URL. W przypadku długich adresów stosunkowo często stosowany był znak przeniesienia części adresu do następnej linii, co skutkowało jego formalną modyfikacją i uzyskaniem komunikatu o niedostępności zasobu. Niekiedy adres łączy był również zakończony kropką. W adresach występowały również zbędne spacje. Zdecydowano się usunąć z adresów URL wszystkie dodatkowe znaki nie wchodzące w skład oryginalnego linku. Wśród błędów w zapisie łączy pojawiały się tzw. literówki, np. zamiast „http:” występował zapis „http;”. Tego rodzaju błędy również zostały skorygowane. Autorzy wykorzystywali również w przypisach bibliograficznych identyfikatory DOI, które w większości przypadków nie miały postaci URL, np. doi:10.1007/978-3-642-04346-8. W celu zapewnienia bezpośredniego dostępu do zasobu zdecydowano się w sposób automatyczny dokonać konwersję tych zapisów do postaci URL poprzez dodanie do identyfikatorów DOI formuły „http://dx.doi.org/”, np. http://dx.doi.org/doi:10.1007/978-3-642-04346-8. Do analizy odesłań wykorzystano aplikację HEADMasterSEO (HEADMasterSEO, 2016). Jest to oprogramowanie, które na podstawie wskazanej listy adresów URL pobiera odpowiedzi serwera w postaci kodów odpowiedzi http. Aplikacja ta przeprowadza pierwszy etap interakcji między klientem a serwerem realizującym dostęp do danego dokumentu, czyli wysyła zapytanie o dostępność zasobu otrzymując ustandaryzowany kod odpowiedzi wskazujący na status jego dostępności. Według dokumentacji protokołu HTTP kody odpowiedzi przyjmują wartość trzycyfrową (od 100 do 511) i można je zorganizować w pięciu grupach (HTTP, 2004):

- (1) Kody informacyjne, np. 110: Przekroczono czas połączenia. Serwer zbyt długo nie odpowiada.
- (2) Kody powodzenia, np. 200: Zawartość żądanego dokumentu.
- (3) Kody przekierowania, np. 301: Trwale przeniesiony – żądany zasób zmienił swój URI i w przyszłości zasób powinien być szukany pod wskazanym nowym adresem.
- (4) Kody błędu aplikacji klienta, np. 404: Nie znaleziono – serwer nie odnalazł zasobu według podanego URL ani niczego co by wskazywało na istnienie takiego zasobu w przeszłości.
- (5) Kody błędu serwera http, np. 500: Wewnętrzny błąd serwera – serwer napotkał niespodziewane trudności, które uniemożliwiły zrealizowanie żądania.

Aplikacja HEADMasterSEO zwracała dla każdego adresu URL kod odpowiedzi serwera HTTP, a w przypadku kodów wskazujących na przekierowania dodatkowo ich liczbę oraz kod odpowiedzi dla ostatecznie ustanowionego adresu URL. W celu określenia wiarygodności uzyskanych danych na losowo dobranej próbie 10% badanych adresów URL przeprowadzono ponowne testy. Wyniki te potwierdziły wcześniej uzyskane kody odpowiedzi.

Badanie przeprowadzono w dniach 9–13 maja 2016 r. korzystając z łącza internetowego o przepustowości 8 Mb/s. z zalecaną przez producenta konfiguracją aplikacji HEADMasterSEO.

4. Wyniki

W tej części artykułu zaprezentowano wyniki badania czterech polskich czasopism z zakresu BIN. Zostały one podzielone na zagadnienia badawcze zgodnie z podziałem przedstawionym w części poświęconej stanowi badań. Wyniki są podstawą do porównania liczby cytowań zasobów internetowych i aktywności cytowanych materiałów w polskich i zagranicznych czasopismach z zakresu BIN oraz przedstawienia rekomendacji związanych z problemem braku stabilności odwołań do zasobów sieciowych.

4.1. Analiza bibliografii załącznikowych i przypisów

W analizowanym zbiorze ok. 37% wszystkich pozycji w bibliografiach załącznikowych i przypisach zawierało odwołania do ich lokalizacji w Internecie. Wskaźnik ten utrzymuje się na poziomie 20–30% dla trzech tytułów czasopism z wyjątkiem *Biuletynu EBIB*, gdzie średnio ponad połowa (58.6%) cytowań, zawiera odwołania do zasobów sieciowych (Tab. 1).

Tabela 1. Liczba i odsetek cytowań z URL w badanych czasopismach

Tytuł czasopisma	Liczba cytowań z URL	Odsetek cytowań z URL	Średnia liczba cytowań z URL na artykuł
Przegląd Biblioteczny	923	31.4	10.3
Bibliotheca Nostra. Śląski Kwartalnik Naukowy	920	23.4	6.4
Biuletyn EBIB	2317	58.6	6.7
Zagadnienia Informatyki – Studia Informacyjne	707	29.0	7.5

Z punktu widzenia pojedynczej publikacji, średnia liczba cytowań dokumentów sieciowych wynosi 7.3 pozycji. Jedynie w odniesieniu do *Przeglądu Bibliotecznego* zaobserwowano wyższą wartość dla tej zmiennej (średnio 10.3). Rozkład wartości dla tego wskaźnika z podziałem na rok publikacji artykułu cytującego prezentuje tabela 2. Największą wartość zmienna ta przyjmuje dla artykułów opublikowanych w 2010 i 2014 r. (odpowiednio 9 oraz 8.1).

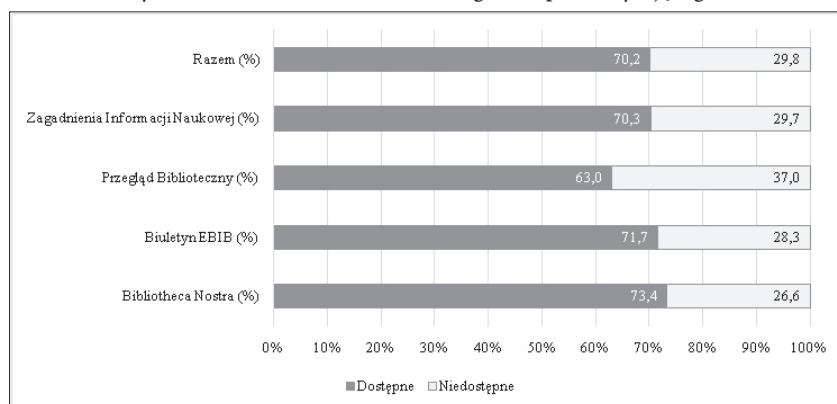
Tabela 2. Średnia liczba cytowań z URL na artykuł wg roku wydania publikacji cytującej

Rok publikacji	Liczba artykułów	Liczba cytowań z URL	Średnia liczba cytowań z URL na artykuł
2009	97	585	6.0
2010	78	705	9.0
2011	90	541	6.0
2012	113	854	7.6
2013	96	724	7.5
2014	104	846	8.1
2015	92	612	6.7
Razem	670	4867	7.3

4.2. Aktualność adresów URL w bibliografiach załącznikowych i przypisach

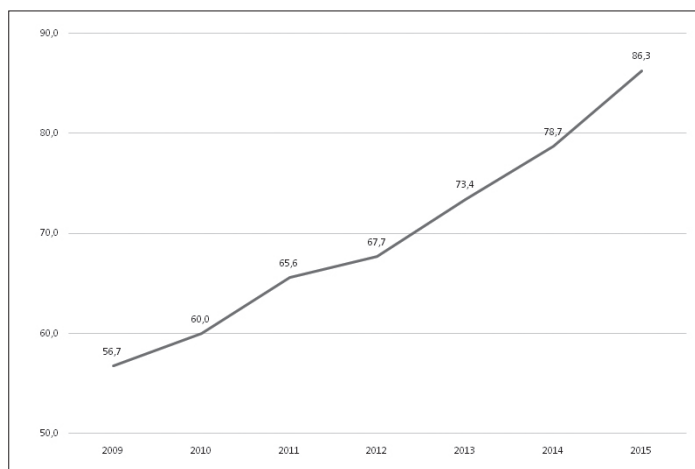
W wyniku przeprowadzonych testów na zbiorze 4593 adresów URL określono, że 70.2% odesłań do zasobów internetowych było aktywnych w maju 2016 r. Największy odsetek nieaktywnych odwołań odnotowano dla *Przeglądu Bibliotecznego* i wynosi on 37%. W pozostałych przypadkach aktualność odesłań waha się w przedziale 70–73% (Rys. 1).

Rys. 1. Aktualność adresów URL wg czasopisma cytującego



Korelując wartość tego wskaźnika z rokiem publikacji artykułu cytującego można odnotować jego sukcesywny wzrost na przestrzeni siedmiu lat. W artykułach z 2009 r. tylko ponad połowa (56.7%) adresów URL w bibliografiach załącznikowych jest obecnie aktualna. W ciągu kolejnych lat odsetek nadal aktywnych odwołań rośnie średnio o 5% i dla artykułów z 2015 r. wynosi on 86.3% (Rys. 2).

Rys. 2. Odsetek aktualnych adresów URL wg roku wydania publikacji cytującej



Szczegółowa analiza zmian wartości wskaźnika dla aktualności adresów URL na przestrzeni siedmiu lat z odniesieniem do źródła publikacji cytującej nie potwierdza jednak

stabilnego wzrostu. O ile w odniesieniu do artykułów opublikowanych w 2015 r. odsetek aktywnych URL jest wyraźnie większy niż dla artykułów z 2009 r., to w okresie pomiędzy tymi datami nie można zauważyć jednolitej tendencji wzrostowej. Co więcej, w każdym z analizowanych tytułów odnotowano przynajmniej jedną taką sytuację, w której odsetek aktywnych URL zmalał w stosunku do jego wartości z roku poprzedniego. Największą stabilność w tym ujęciu reprezentuje *Biuletyn EBIB*. Uwagę zwraca stosunkowo niski odsetek aktywnych adresów URL zamieszczanych w *Przeglądzie Bibliotecznym* w latach 2009–2010 (Tab. 3). Nie udało się jednak ustalić przyczyn tego zjawiska.

Tabela 3. Odsetek aktualnych adresów URL dla poszczególnych tytułów czasopism w latach 2009–2015

	2009	2010	2011	2012	2013	2014	2015
Bibliotheca Nostra. Śląski Kwartalnik Naukowy	70.0	65.4	68.9	66.4	79.9	75.3	86.7
Biuletyn EBIB	60.2	64.7	63.5	69.4	76.0	79.0	85.4
Przegląd Biblioteczny	39.7	35.7	67.6	69.3	62.6	78.5	82.4
Zagadnienia Informacji Naukowej – Studia Informacyjne	52.5	71.5	60.6	63.0	70.5	82.7	93.7

4.3. Odpowiedzi serwera

Analizując bezpośrednie odpowiedzi serwerów na żądanie pobrania nagłówka HTTP przez aplikację HEADMasterSEO (Tab. 4) można zaobserwować, że tylko w ok. 44% wskazań zasoby internetowe były bezpośrednio dostępne pod wskazanym adresem. Jednocześnie odnotowano duży odsetek przekierowań (32.3%). W celu uzyskania pełnego obrazu poziomu dostępności do cytowanych zasobów sieciowych szczegółowej analizie poddano grupę adresów URL, dla których narzędzie badawcze zwróciło kody nagłówków z grupy 3XX. Wyniki tych badań wskazują, że w 79% przypadków przekierowania doprowadziły do uzyskania prawidłowego rezultatu, czyli zgłoszenia kodu 200 – zasób sieciowy jest dostępny. Tym samym średni rzeczywisty poziom dostępności dla analizowanego zbioru odwołań wynosi, tak jak wspomniano wcześniej, 70.2%.

Tabela 4. Grupy kodów http uzyskane dla badanych adresów URL

Kod odpowiedzi	Odsetek wskazań
2XX	44.6
3XX	32.3
4XX	18.0
5XX	1.0
Pozostałe błędy	4.1

W zbiorze adresów URL, dla których odnotowano kod odpowiedzi z grupy 3XX występują łańcuchy przekierowań o wartościach od 1 do 10. Jednokrotne przekierowania

dotyczą 72% adresów z tej puli, dwukrotne – 19%, trzykrotne – 6%, czterokrotne – 1.6%, a pozostałe uzyskały wartości mniejsze niż 1%.

Spośród odpowiedzi odnoszących się do nieaktywnych adresów URL największy odsetek stanowił kod 404 (75.2%) oznaczający, że serwer nie odnalazł zasobu według podanego URL. Na drugim miejscu (16.4%) pod względem liczby wystąpień znalazła się grupa odpowiedzi nie w postaci kodu HTTP, lecz odpowiedzi aplikacji związanej z brakiem możliwości nawiązania połączenia (np. przekroczony czas oczekiwania, błąd połączenia). Pozostałe odpowiedzi pojawiały się bardzo rzadko, np. kody 403 (4.2%), 500 (3.6%) oraz 410 (2.7%).

4.4. Wskaźnik połowicznej dezaktywacji łączy (*half-life*)

Kolejnym etapem analizy uzyskanych danych badawczych było obliczenie wskaźnika połowicznej dezaktywacji łączy. Określa on okres liczony w latach, po którym połowa identyfikatorów sieciowych przestaje być aktywna. W celu obliczenia wskaźnika posłużono się następującym wzorem (Dimitrova & Bugeja, 2007, 820):

$$t_h = \frac{t \ln(0,5)}{\ln W_{(t)} - \ln W_{(0)}}$$

Zmienna t_h oznacza czas połowicznej dezaktywacji łączy, $W_{(0)}$ oznacza liczbę aktywnych łączy w chwili publikacji artykułu, a $W_{(t)}$ – liczbę aktywnych łączy w późniejszym punkcie czasu t . Dla całego zbioru danych badawczych, a więc w odniesieniu do czterech analizowanych tytułów czasopism, wskaźnik połowicznej dezaktywacji łączy przyjął wartość 7.39. Tabela 5 przedstawia rozkład jego wartości dla poszczególnych tytułów czasopism w odniesieniu do roku publikacji artykułów cytujących.

Tabela 5. Rozkład wartości wskaźnika połowicznej dezaktywacji łączy (*half-life*)

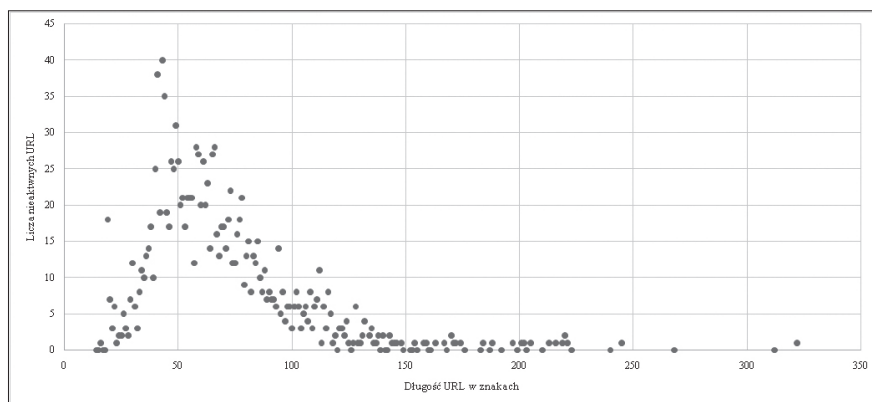
	2009	2010	2011	2012	2013	2014	2015	Średnia
Bibliotheca Nostra. Śląski Kwartalnik Naukowy	13.60	9.79	9.32	6.76	9.25	4.90	4.87	8.36
Biuletyn EBIB	9.55	9.55	7.64	7.58	7.59	5.89	4.39	7.46
Przegląd Biblioteczny	5.26	4.04	8.86	7.56	4.44	5.74	3.58	5.64
Zagadnienia Informacji Naukowej – Studia Informacyjne	7.53	12.42	6.92	5.99	5.96	7.28	10.60	8.10

Wartość wskaźnika jest różna dla poszczególnych czasopism na przestrzeni badanego okresu. Stały trend jest widoczny jedynie w przypadku łączy zamieszczonych w *Biuletynie EBIB*. Najniższą wartość współczynnik przyjął w przypadku *Przeglądu Bibliotecznego* (5.64). Jest to jedyne czasopismo spośród analizowanych w artykule, w którym połowa cytowanych łączy dla wybranych lat była nieaktywna w momencie przeprowadzania badania. Dotyczyło to lat 2009 (5.26) i 2010 (4.04).

4.5. Budowa adresów URL a poziom dostępności

Omawiając stan badań wspomniano, że długość adresu URL liczona w znakach może być rozważana jako jeden z czynników wpływających na aktywność odesłania. Średnia długość adresu URL w analizowanym zbiorze wyniosła 60 znaków. Dla aktywnych odwołań wskaźnik ten przyjął wartość 57 znaków, a dla nieaktywnych 67. W wyniku analizy danych nie stwierdzono liniowego wzrostu liczby nieaktywnych łączy, czyli zależności między długością adresu URL a dostępnością zasobu sieciowego (Rys. 3).

Rys. 3. Zależność między długością adresu URL a jego dostępnością



Źródło: badania własne.

Analiza tego zjawiska z zastosowaniem organizacji danych w przedziałach znakowych odpowiadających długości adresu URL oraz odsetka ich dostępności również nie potwierdza tej hipotezy (Tab. 6). Widać jednak, że największy odsetek odwołań to adresy URL o długości od 41 do 60 znaków i to w tej grupie odnotowano największą liczbę nieaktywnych łączy.

Tabela 6. Zależność między długością adresu URL a jego dostępnością

Długość URL	Liczba URL	Odsetek URL	Odsetek nieaktywnych URL
1–20	170	3.70	1.90
21–40	952	20.73	11.70
41–60	1645	35.82	35.41
61–80	934	20.34	26.04
81–100	484	10.54	12.66
101–120	245	5.33	7.24
121–140	96	2.09	2.85
>141	67	1.46	2.19

Nawiązując do hipotezy wspomnianej w pierwszej części tekstu, wskazującej na zależność poziomu dostępności zasobu od głębokości adresu URL, czyli liczby katalogów w łączy, nie stwierdzono takiej zależności w odniesieniu do zgromadzonych danych (Tab. 7).

Tabela 7. Zależność między głębokością adresu URL a jego dostępnością

Głębokość URL	Liczba URL	Liczba nieaktywnych URL	Odszetek nieaktywnych URL
0	1	0	0
1	1136	225	19.8
2	1208	352	29.1
3	977	342	35.0
4	624	217	34.8
5	345	140	40.6
6	179	54	30.2
7	81	22	27.2
8	16	8	50.0
9	19	6	31.6
10	4	0	0
11	2	1	50.0
13	1	0	0

Co prawda, w przypadku pięciopoziomowych adresów URL połowa z odwołań była nieaktywna, lecz tendencja ta nie utrzymywała się na wyższych poziomach, a liczba takich adresów nie była znacząca. Ponad 70% adresów URL w zbiorze wskazywało na dosyć płytką strukturę odwołań (do trzech podkatalogów) i w tej grupie można zaobserwować wzrost poziomu niedostępności zasobu wraz ze zwiększającą się głębokością odwołania.

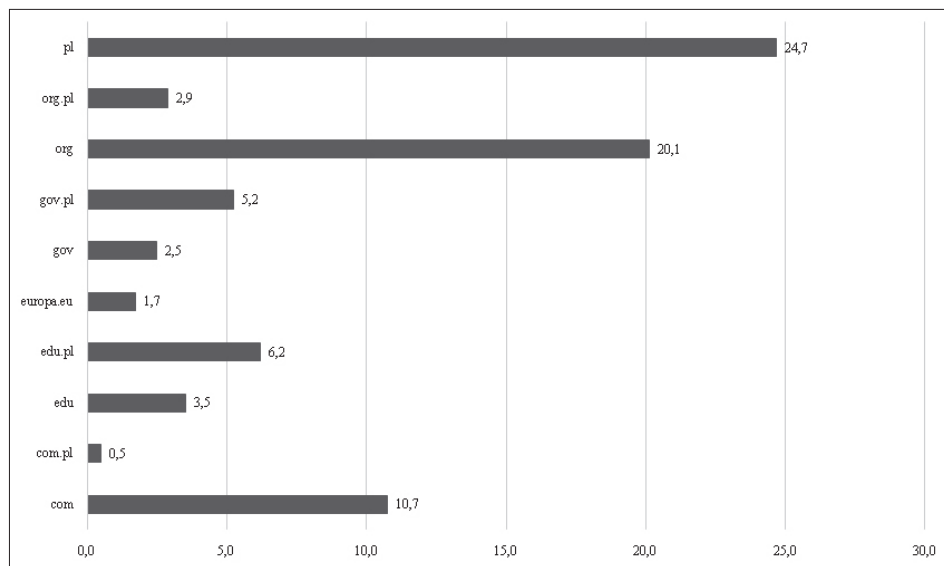
4.6. Pochodzenie cytowanych zasobów sieciowych oraz jego wpływ na poziom ich dostępności

Celem prac badawczych, oprócz określenia poziomu dostępności cytowanych zasobów internetowych, było również naszkicowanie ogólnego obrazu tego, co cytują autorzy publikacji. W tym celu z puli adresów URL automatycznie wyodrębniono typy domen internetowych oraz nazwy domenowe. W pierwszym przypadku zdecydowano zastosować podstawową typologię domen internetowych z dodatkowym uwzględnieniem rozszerzenia *pl* jako wskazania na rejestrację domeny w Polsce oraz *europa.eu* jako wskazania na rejestrację domeny w ramach usług informacyjnych Unii Europejskiej (Rys. 4).

Z punktu widzenia typu domeny największy odsetek adresów URL wskazuje na zasoby cyfrowe udostępniane przez organizacje pozarządowe i non profit (*org* – 20.1%). Jedynie niecałe 3% adresów URL wskazywało na tego typu źródła zarejestrowane w Polsce (*org.pl*). Ponad 11% adresów URL było związanych z podmiotami komercyjnymi, z czego tylko 0.5% miało zarejestrowano w domenie *com.pl*. Warto podkreślić, że domena ta może być rejestrowana nie tylko przez firmy, ale również przez osoby prywatne. Bardzo podobny odsetek wystąpień, co domena *com*, miały również strony należące do instytucji edukacyjnych, z tym

że domena ogólna *edu* (3.5%) miała mniej wystąpień niż jej polska odmiana *edu.pl* (6.2%). Około 8% adresów przekierowywało do serwisów informacyjnych instytucji rządowych (*gov* i *gov.pl*). Uwzględniając narodowy charakter domeny, prawie jedna czwarta adresów zawierała wyłącznie rozszerzenie *pl*. Nieuwzględnione na rysunku 4 pozostałe typy domen były reprezentowane przez ok. 22% adresów URL ze zbioru badawczego.

Rys. 4. Typy domen (%)



Analiza danych pod kątem nazw domenowych, występujących w adresach URL, nie doprowadziła do wskazania dominujących sieciowych źródeł informacji występujących w przypisach i bibliografiach załącznikowych publikacji cytujących. W tabeli 8 przedstawiono jednak dziesięć internetowych źródeł informacji o największym odsetku cytowań.

Tabela 8. Nazwy domenowe w adresach URL (%)

Domena	Odsetek wskazań
http://www.ebib.info/	2.3
http://goo.gl/	2.2
http://pl.wikipedia.org/	1.9
http://ur.oss.wroc.pl/	1.6
http://www.ifla.org/	1.1
http://dx.doi.org/	1.0
http://www.ala.org/	0.8
http://eprints.rclis.org/	0.7
http://www.bn.org.pl/	0.7
http://www.nauka.gov.pl/	0.7

W analizowanym zbiorze największą liczbę wskazań odnotowano do zasobów pochodzących z serwisu internetowego EBIB. Serwis ten jest notabene niedostępny pod tym adresem, a usługodawca nie zapewnił przekierowania do aktualnej lokalizacji zasobów, co spowodowało, że adresy URL zostały oznaczone przez narzędzie badawcze jako niedostępne. Odnotowano również 99 adresów URL, które wskazywały na wykorzystanie serwisu Google umożliwiającego utworzenie tymczasowego aliasu dla długiego adresu URL. Niestety, 21% z nich było aktualnie nieaktywnych. Na ten problem zwracają uwagę m.in. Klein et al. (2014). W bibliografiach załącznikowych i w przypisach widoczne jest wykorzystanie zasobów polskojęzycznej Wikipedii. W zbiorze badawczym odnotowano 88 odwołań do tego źródła, co stanowi niecałe 2% wszystkich wskazań. Autorzy publikacji w niewielkim stopniu wykorzystują stałe identyfikatory DOI. W zbiorze badawczym odnotowano tylko 44 adresy URL, które wykorzystywały ten system. Trudno jednak, wyłącznie na podstawie tej informacji, stawiać wnioski bez znajomości danych wskazujących na liczbę zarejestrowanych DOI dla wszystkich cytowań w zbiorze danych badawczych.

Interesujące wyniki uzyskano zestawiając typy domen, które reprezentowały zgromadzone adresy URL z poziomem ich aktualności (Tab. 9). Największą grupę odwołań (24.7%) stanowiły adresy URL zakwalifikowane do typu domeny *pl*. Jedna czwarta z nich okazała się być nieaktywna w okresie, w którym prowadzono badanie. Podobną zależność odnotowano w odniesieniu do typu *com*. W przypadku zasobów sieciowych opublikowanych w serwisach instytucji edukacyjnych, jedna trzecia adresów URL była nieaktywna. W przypadku polskich zasobów edukacyjnych odsetek ten wynosi 21.8%. Wysoką liczbę nieaktywnych adresów URL odnotowano dla instytucji rządowych. W przypadku Polski, to prawie połowa odesłań. Również ponad połowa adresów URL do zasobów instytucji Unii Europejskiej okazała się być nieaktywna.

Tabela 9. Pochodzenie cytowanych zasób a stopień ich dostępności

Typ domeny	Dostępne (%)	Niedostępne (%)	Udział (%)
com	77.3	22.7	10.7
com.pl	69.6	30.4	0.5
edu	69.8	30.2	3.5
edu.pl	78.2	21.8	6.2
europa.eu	46.8	53.2	1.7
gov	69.3	30.7	2.5
gov.pl	50.6	49.4	5.2
org	75.5	24.5	20.1
org.pl	81.8	18.2	2.9
pl	74.3	25.7	24.7
pozostałe	60.4	39.6	21.9

Dopełnieniem badań nad dostępnością odwołań do zasobów sieciowych w bibliografiach załącznikowych i przypisach była próba weryfikacji pośredniej dostępności do zasobów zakwalifikowanych jako niedostępne. W tym celu wykorzystano cyfrowe archiwum zasobów

Internetu Wayback Machine (Wayback Machine, b.d.), które jest tworzone w ramach projektu Internet Archive. Na potrzeby badań wykorzystano interfejs programistyczny tej kolekcji Wayback Machine API (Application Programming Interface) (Wayback Machine API, b.d.), który posłużył do weryfikacji dostępności adresów URL. Uzyskane dane wskazują, że 53% zasobów sieciowych zidentyfikowanych w badaniu jako niedostępne posiada kopię ich zawartości w bazie Wayback Machine. Oznacza to, że faktyczny poziom dostępności tych odwołań w całym zbiorze badawczym jest znacznie wyższy. Niestety, dostęp do tych źródeł internetowych wymaga od osób zainteresowanych manualnego ich odnalezienia w zasobach Internetu.

4.7. Zmiana treści cytowanych zasobów internetowych

Istotnym problemem, obok nieaktywności łączy do cytowanych dokumentów, jest kwestia zmiany treści tych zasobów. W badaniu podjęto próbę określenia odsetka cytowanych zasobów internetowych, których treść uległa modyfikacji od czasu odwołania się do nich w analizowanym zbiorze. Analizę przeprowadzono dla każdego czasopisma osobno. Pod uwagę wzięto wyłącznie te łącza, w przypadku których ostateczny kod odpowiedzi HTTP przyjął wartość 200. W celu oszacowania próby wykorzystano wzór (Jabłońska & Sobieraj, 2013, 35):

$$n_b = \frac{N}{1 + \frac{4d^2(N-1)}{Z^2}}$$

Zmienna N oznacza wielkość populacji, wielkość d oznacza zakładany błąd (5%), a wielkość Z wartość standardową 1.96 dla poziomu istotności $p = 0.05$. W analizie zmiany treści przywoływanych zasobów wykorzystano informacje bibliograficzne zamieszczone w przypisach bądź w bibliografiach załącznikowych. Jako zasoby o zmienionej treści oznaczano wyłącznie dokumenty, w przypadku których nie było żadnych wątpliwości, że taka zmiana nastąpiła (Tab. 10). Oznacza to, że wyliczony odsetek zmienionych treściowo zasobów to wartość minimalna, która jest zapewne nieco mniejsza od faktycznie występującej.

Tabela 10. Zmiana treści cytowanych zasobów w poszczególnych czasopismach

Czasopismo	Odsetek zasobów ze zmienioną treścią
Bibliotheca Nostra. Śląski Kwartalnik Naukowy	11.16%
Biuletyn EBIB	9.71%
Przegląd Biblioteczny	15.32%
Zagadnienia Informacji Naukowej – Studia Informacyjne	14.15%

Średni odsetek zmienionych treściowo zasobów wyniósł 12.28%. Wyniki powyżej tej średniej uzyskano dla *Przeglądu Bibliotecznego* oraz *Zagadnień Informacji Naukowej*. Dla porównania w badaniu Mardaniego i Sangariego uzyskano wynik na poziomie 11% (Mardani & Sangari, 2013, 35–36).

5. Dyskusja

W artykule zaprezentowano analizę pozycji z bibliografii załącznikowych oraz przypisów zawierających odesłania do zasobów internetowych. Jako źródło danych badawczych wybrano cztery polskie czasopisma z zakresu BIN. Trzy z nich znajdują się na liście czasopism punktowanych opublikowanej przez Ministerstwo Nauki i Szkolnictwa Wyższego, jedno zaś jest internetowym czasopismem fachowym. Duża część poczynionych ustaleń jest zbieżna z wynikami badań przytoczonymi w części poświęconej stanowi badań, co potwierdza występowanie pewnych ogólnych trendów w zakresie aktywności łączy do cytowanych zasobów internetowych.

W czasopismach uwzględnionych w badaniu ok. 37% wszystkich pozycji w bibliografiach załącznikowych i przypisach zawierało łączy do zasobów internetowych. Średnia ta jest wyższa od uzyskanej w innych badaniach. Te niestety obejmują wcześniejsze lata publikacji, co w zestawieniu z prezentowanymi wynikami badań nie pozwala na jednoznaczne wyciąganie wniosków, ale może stanowić pewien punkt odniesienia. Ali Sadat-Moosavi, Alireza Isfandiyari-Moghaddam oraz Oranus Tajeddini obliczyli, że średni odsetek cytowanych zasobów sieciowych w czterech czasopismach dotyczących BIN w latach 2005–2008 wyniósł 24% (Sadat-Moosavi et al., 2012, 184), a Kumar, Kumar oraz Parameshwarappa określili jego poziom na 23.81% (Kumar et al., 2015, 61) dla publikacji z lat 2008–2012. W ostatnim przytoczonym badaniu zaobserwowano spadek odsetka cytowanych zasobów internetowych. Procent cytowań z URL w czasopiśmie *Bibliotheca Nostra. Śląski Kwartalnik Naukowy* osiągnął najbliższą wartość w stosunku do średniej wyliczonej w innych badaniach, a internetowe czasopismo fachowe *Biuletyn EBIB* znacznie ją przekroczyło. Trudno jest jednak wskazać pewne przyczyny odmiennego stosunku pomiędzy powoływaniem się na materiały drukowane, a odwoływaniem się do zasobów sieciowych w polskich czasopismach z zakresu BIN. Może się to wiązać m.in. z czasem pochodzenia materiału badawczego. Należy podkreślić, że zwiększa to zagrożenie niedostępności dokumentów przytaczanych przez autorów artykułów.

Średnia liczba cytowań zasobów internetowych na artykuł dla polskich czasopism jest zazwyczaj wyższa (7.3) niż średnia dla ich zagranicznych odpowiedników. W jednym z badań czasopism z zakresu BIN (*The Electronic Library, Program*) z lat 2008–2012 wyliczono średnią liczbę cytowanych zasobów sieciowych na artykuł w wysokości 6.10 (Kumar et al., 2015, 61). Niekiedy jednak wynik jest znacznie wyższy, jak dla internetowego czasopisma *Ariadne* (2010–2012), w przypadku którego średnia wyniosła aż 16.21 (Gul et al., 2014, 31). Wynik ten jest rozbieżny z uzyskanym dla *Biuletynu EBIB*. Wysoka średnia w przypadku *Przeglądu Bibliotecznego* nie wiąże się z ogólnie wysoką liczbą pozycji bibliograficznych przywoływanych w tym czasopiśmie. Zaprezentowane wyniki są często rozbieżne, co wskazuje na ścisłą zależność omawianego wskaźnika od charakteru czasopisma.

W wyniku analiz określono, że 70.2% odesłań do zasobów internetowych było aktywnych w maju 2016 r. Jest to stosunkowo wysoki wskaźnik w porównaniu z uzyskanymi w innych badaniach czasopism z dziedziny BIN. Najbardziej zbliżony wynik uzyskali Goh i Ng – 69% (Goh & Ng, 2007, 19). Wskazuje to jednak, że zgodnie z oczekiwaniami niedostępność cytowanych zasobów internetowych jest istotnym problemem w przypadku polskich czasopism z dziedziny BIN. Problem ten nie jest równie widoczny we wszystkich badanych czasopismach. Najmniejszy odsetek aktywnych odesłań odnotowano dla *Przeglądu*

Bibliotecznego – 63%, co pokazuje wagę omawianego zagadnienia w przypadku tego periodyku. Problem występuje jednak również w innych czasopismach. Należy podkreślić, że redakcje czasopism mają tylko ograniczone możliwości wpływania na zwiększenie dostępności cytowanych zasobów. We wskazówkach dla autorów w trzech analizowanych czasopismach naukowych znajdują się dokładne zalecenia dotyczące sposobu cytowania dokumentów sieciowych. Poziom dostępności dokumentów internetowych można poprawić sprawdzając wszystkie linki przed ich użyciem i opublikowaniem w czasopiśmie, co było już podkreślane przez niektórych autorów (zob. np. Dimitrova & Bugeja, 2007, 822; Prithviraj & Kumar, 2014, 45). W przypadku *Biuletynu EBIB* problemem jest zaobserwowana w czasie badań duża zmienność sposobu cytowania zasobów internetowych. Niekiedy nawet w ramach jednego numeru stosowane były różne rozwiązania: przypisy, bibliografia załącznikowa bądź zamieszczanie linków w tekście artykułu. Z poruszonymi zagadnieniami ściśle wiąże się wartość współczynnika połowicznej dezaktywacji łączy. Średnia dla analizowanych czasopism jest stosunkowo wysoka i wynosi 7.39 lat. Wartość tego wskaźnika jest wyraźnie niższa dla *Przeglądu Bibliotecznego* (5.64). Jest to jednak wartość wyższa, niż uzyskiwana często w innych badaniach. Podkreślono już, że wysokie wyniki uzyskano dla czasopism internetowych *D-Lib Magazine* – 10 lat (McCown et al., 2005, 5) oraz *Information Research* – 14.04 lat (Moghaddam et al., 2010, 71–72). W przypadku polskiego czasopisma *Biuletyn EBIB* wynik jest dużo niższy.

Najczęściej występującą odpowiedzią HTTP w przypadku nieaktywnych łączy był kod 404 stanowiący 75.2% odpowiedzi. Jest to wynik podobny do uzyskanych w innych badaniach. Odmienność polega na dużo większym odsetku tego typu odpowiedzi w prezentowanym badaniu polskich czasopism. Inne kody stanowiły niewielki odsetek wszystkich błędnych odpowiedzi. Istotnym problemem pozostaje zatem zmiana układu serwisów i przenoszenie zasobów.

W badaniach uwzględniono również hipotezę mówiącą, że budowa adresu URL może mieć wpływ na jego aktualność. Założenie to przyjęto na podstawie wyników stanu badań, gdzie badacze (np. Spinellis, 2003; Goh & Ng, 2007; Prithviraj & Kumar, 2014; Kumar et al., 2015) zaobserwowali pewne zależności w tym kontekście w odniesieniu do długości adresu URL oraz jego głębokości. W trakcie badania nie stwierdzono zależności między długością adresu URL a dostępnością zasobu. Liczba znaków dla zasobów niedostępnych jest co prawda nieco większa, jednak nie jest to czynnik, który z dużym prawdopodobieństwem pozwala określić stabilność łączy. Podobnie nie stwierdzono zależności pomiędzy poziomem dostępności zasobu a głębokością adresu URL.

W przeprowadzonym badaniu popularne rodzaje domen nie różnią się od wyników wcześniej przeprowadzonych analiz. Najczęściej cytowaną domeną była domena *org*, a następnie *com*, *edu* oraz *gov*. Hipotezy dotyczące popularności poszczególnych domen zostały przedstawione w części poświęconej stanowi badań. Łączy z najpopularniejszej domeny *org* należały do jednych z najbardziej trwałych. Z kolei do najmniej stabilnych linków należały, oprócz łączy zawierających nazwę domeny wskazującej na serwisy Unii Europejskiej *europa.eu*, te należące do polskich instytucji rządowych. Wskazuje to na dużą zmienność tych stron. Być może wśród cytowanych dokumentów dużą liczbę stanowiły różnego rodzaju ogłoszenia, które z racji swego tymczasowego charakteru są usuwane ze stron jako nieaktualne. W celu potwierdzenia tej hipotezy należałoby dokładnie przeanalizować rodzaje cytowanych zasobów ze stron rządowych.

W badaniu podjęto również próbę analizy możliwości odszukania zasobów zakwalifikowanych jako niedostępne wykorzystując internetowe archiwum Wayback Machine. Za jego pomocą udało się uzyskać dostęp do 53% wcześniej niedostępnych dokumentów. Wskazuje to na możliwość manualnego odnalezienia niedostępnych pierwotnie zasobów, co jest jednak dość kłopotliwe.

Oprócz niedostępności zasobów pod łączami zamieszczonymi w bibliografiach załącznikowych, problemem wydaje się być również zmiana treści cytowanych dokumentów sieciowych. W badaniu określono poziom tego zjawiska na 12.28%. Z jednej strony odsetek ten nie jest wysoki, jednak w połączeniu z liczbą niedostępnych zasobów, powoduje poważne ograniczenia w dostępie do cytowanych materiałów zmniejszając tym samym możliwość późniejszej analizy prac naukowych.

6. Zakończenie

Na koniec warto zastanowić się nad możliwymi do wdrożenia rozwiązaniami, które zapewniłyby większą dostępność cytowanych zasobów internetowych. Ciągłe aktualne zostaje zalecenie dla redakcji czasopism, które powinny prowadzić spójną politykę odnośnie cytowania materiałów sieciowych. Wspomniano już o tym, że redakcje czasopism powinny sprawdzać aktywność łączy przed publikacją artykułów. Tymczasowe rozwiązanie może opierać się na istniejących, amerykańskich serwisach archiwizacji Internetu, takich jak Internet Archive, WebCite lub Perma. Na początku artykułu wspomniano o ograniczeniach związanych z korzystaniem z tych narzędzi. Wydaje się, że mimo ich występowania redakcje czasopism mogłyby umożliwić zapisywanie obok oryginalnego adresu URL cytowanego zasobu, łączy do serwisu archiwizacyjnego. Na przykład styl cytowania Bluebook umożliwia zamieszczanie linków do serwisu perma.cc (Kribble, 2015). Bardziej trwałe rozwiązanie powinno opierać się na stworzeniu polskiego serwisu archiwizującego cytowane zasoby, czy szerzej dokumenty ważne dla polskiej nauki i kultury. W celu jego stworzenia można by się posłużyć istniejącymi zasobami, np. kod Perma jest dostępny w serwisie GitHub (GitHub, b.d.). Obsługą takiego serwisu powinna zająć się instytucja z zapewnionym stałym finansowaniem. Wśród bibliotek najlepiej przygotowana do tego zadania wydaje się być Biblioteka Narodowa, dysponująca dobrym zapleczem technicznym. Inne rozwiązanie może polegać na oparciu archiwizacji na sieci instytucji, podobnie jak zrobiono to w Stanach Zjednoczonych w przypadku serwisu Perma. W Polsce największa liczba bibliotek naukowych i akademickich jest skupiona wokół Narodowego Uniwersalnego Katalogu Centralnego (NUKAT). Wydaje się, że Centrum NUKAT zarządzające projektem dysponuje sporym doświadczeniem w zakresie organizacji współpracy między różnymi bibliotekami. Inną instytucją, która mogłaby podjąć się tego zadania może być Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy, operator Polskiej Bibliografii Naukowej. Niezależnie od przyjętego rozwiązania organizacyjnego wydaje się, że stworzenie polskiego serwisu archiwizacyjnego jest sprawą istotną i wymagającą rozwiązania w najbliższej przyszłości. Dalsze badania cytowanych zasobów internetowych mogłyby objąć polskie czasopisma z innych dziedzin wiedzy. Wydaje się jednak, że ważniejszą sprawą jest opracowanie metodyki archiwizacji zasobów i wdrożenie rozwiązania.

Bibliografia

- archive.is (b.d.). [online], [09.06.2016], <http://archive.is/>
- Carnevale, R. J.; Aronsky, D. (2007). The life and death of URLs in five biomedical informatics journals [online]. *International Journal of Medical Informatics*, 76(4), 269–273, [09.06.2016], <http://doi.org/10.1016/j.ijmedinf.2005.12.001>
- Casserly, M. F.; Bird, J. E. (2003). Web citation availability: Analysis and implications for scholarship [online]. *College and Research Libraries*, 64(4), 300–317, [09.06.2016], <http://doi.org/10.5860/crl.64.4.300>
- Dimitrova, D. V.; Bugeja, M. (2007). The half-life of internet references cited in communication journals [online]. *New Media & Society*, 9(5), 811–826, [09.06.2016], <http://doi.org/10.1177/1461444807081226>
- GitHub (b.d.). GitHub – harvard-lil/perma: Indelible links [online]. [09.06.2016], <http://github.com/harvard-lil/perma>
- Goh, D. H.-L.; Ng, P. K. (2007). Link decay in leading information science journals [online]. *Journal of the American Society for Information Science and Technology*, 58(1), 15–24, [09.06.2016], <http://doi.org/10.1002/asi.20513>
- Gul, S.; Mahajan, I.; Ali, A. (2014). The growth and decay of URLs citation: A case of an online library & information science journal [online]. *Malaysian Journal of Library and Information Science*, 19(3), 27–39, [09.06.2016], <http://ejum.fsktm.um.edu.my/ArticleInformation.aspx?ArticleID=1502>
- Habibzadeh, P. (2013). Decay of references to Web sites in articles published in general medical journals: mainstream vs small journals [online]. *Applied Clinical Informatics*, 4(4), 455–464, [09.06.2016], <http://doi.org/10.4338/ACI-2013-07-RA-0055>
- HEADMasterSEO. (2016). Wersja 1.07. [online]. [09.06.2016], <http://headmasterseo.com/>
- HTTP (2004). Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content [online], [09.06.2016], <http://tools.ietf.org/html/rfc7231>
- Internet Archive (b.d.). Internet Archive Frequently Asked Questions [online], [09.06.2016], <http://archive.org/about/faqs.php>
- Jabłońska, K.; Sobieraj, A. (2013). Metodyka dobierania próby badawczej w naukach społecznych [online]. *Bezpieczeństwo i Technika Pożarnicza*. 32(4), 31–36, [09.06.2016], <http://doi.org/10.12845/bitp.32.4.2013.3>
- Klein, M.; Van de Sompel, H.; Sanderson, R.; Shankar, H.; Balakireva, L.; Zhou, K.; Tobin, R. (2014). Scholarly context not found: one in five articles suffers from reference rot [online]. *PLoS One*, 9(12), [09.06.2016], <http://doi.org/10.1371/journal.pone.0115253>
- Koehler, W. (2004). A longitudinal study of Web pages continued: a consideration of document persistence [online]. *Information Research: An International Electronic Journal*, 9(2), [09.06.2016], <http://www.informationr.net/ir/9-2/paper174.html>
- Kribble, M. (2015). 20th edition of the Bluebook: including Perma.cc! [online], [09.06.2016], <http://etseq.law.harvard.edu/2015/06/20th-edition-of-the-bluebook-including-perma-cc/>
- Kumar, D. V.; Kumar, B. T. S.; Parameshwarappa, D. R. (2015). URLs Link Rot: Implications for Electronic Publishing [online]. *World Digital Libraries*, 8(1), 59–65, [09.06.2016], <http://doi.org/10.18329/09757597/2015/8105>
- Mardani, A. H.; Sangari, M. (2013). An analysis of the availability and persistence of web citations in Iranian LIS journals [online]. *International Journal of Information Science and Management*, 11(1), 29–42, [09.06.2016], <http://ijism.ricest.ac.ir/index.php/ijism/article/view/202>
- McCown, F.; Sheffan, C.; Nelson, M. L.; Bollen, J. (2005). The Availability and Persistence of Web References in D-Lib Magazine [online]. In: *5th International Web Archiving Workshop (IWA05)*, [09.06.2016], <http://arxiv.org/abs/cs/0511077v1>
- Moghaddam, A. I.; Saberi, M. K.; Esmaeel, S. M. (2010). Availability and Half-life of Web References Cited in Information Research Journal: A Citation Study [online]. *International Journal of*

- Information Science and Management (IJISM)*, 8(2), 57–75, [09.06.2016], <http://ijism.ricest.ac.ir/index.php/ijism/article/view/49>
- Oguz, F.; Koehler, W. (2016). URL decay at year 20: A research note [online]. *Journal of the Association for Information Science and Technology*, 67(2), 477–479, [09.06.2016], <http://doi.org/10.1002/asi.23561>
- Perma.cc (b.d.). [online], [09.06.2016], <http://perma.cc/>
- Perma.cc Docs (b.d.). [online], [09.06.2016], <http://perma.cc/docs>
- Prithviraj, K. R.; Kumar, B. T. S. (2014). Corrosion of URLs: Implications for electronic publishing [online]. *IFLA Journal*, 40(1), 35–47, [09.06.2016], <http://doi.org/10.1177/0340035214526529>
- Sadat-Moosavi, A.; Isfandyari-Moghaddam, A.; Tajeddini, O. (2012). Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals [online]. *Aslib Proceedings*, 64(2), 178–192, [09.06.2016], <http://doi.org/10.1108/00012531211215196>
- Sampath Kumar, B. T.; Prithvi Raj, K. R. (2012). Availability and persistence of web citations in Indian LIS literature [online]. *The Electronic Library*, 30(1), 19–32, [09.06.2016], <http://doi.org/10.1108/02640471211204042>
- Spinellis, D. (2003). The decay and failures of web references [online]. *Communications of the ACM*, 46(1), 71–77, [09.06.2016], <http://doi.org/10.1145/602421.602422>
- Van de Sompel, H.; Klein, M.; Jones, S. M. (2016). Persistent URLs Must Be Used To Be Persistent [online]. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, 119–120, [09.06.2016], <http://arxiv.org/abs/1602.09102v1>
- Wayback Machine (b.d.). [online], [09.06.2016], <http://archive.org/web>
- Wayback Machine APIs (b.d.). [online], [09.06.2016], http://archive.org/help/wayback_api.php
- WebCite (b.d.). [online]. [09.06.2016], <http://www.webcitation.org/>
- Wu, Z. (2008). An empirical study of the accessibility of web references in two Chinese academic journals [online]. *Scientometrics*, 78(3), 481–503, [09.06.2016], <http://doi.org/10.1007/s11192-007-1951-1>
- Yang, S.; Han, R.; Ding, J.; Song, Y. (2012). The distribution of Web citations [online]. *Information Processing and Management*, 48, 779–790, [09.06.2016], <http://doi.org/10.1016/j.ipm.2011.10.002>
- Zhou, K.; Grover, C.; Tobin, R.; Klein, M. (2015). No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving [online]. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (2015th-June ed., Vol. 2015-June, 233–236), [09.06.2016], <http://doi.org/10.1145/2756406.2756940>
- Zittrain, J.; Albert, K.; Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations [online]. *Legal Information Management*, 14(02), 88–99, [09.06.2016], <http://doi.org/10.1017/S1472669614000255>
-

Web Citations in Polish Library and Information Science Journals: the Analysis of URL Validity

Abstract

Purpose/Thesis: This paper is intended to provide a comprehensive literature review on the issue of link rot, that is, the inaccessibility of Web citations in academic journals and describe this problem based on articles published in four Polish journals in the field of library and information science.

Approach/Research methods: Using critical literature review, the authors identified main areas of interest related to the problem of link rot in scientific journals, as well as key methodological approaches.

The subject of empirical studies were web citations in research papers published in four Polish library and information science journals from 2009 to 2015: *Bibliotheca Nostra*, *Śląski Kwartalnik Naukowy*, *Biuletyn EBIB*, *Przegląd Biblioteczny*, *Zagadnienia Informatyki Naukowej - Studia Informacyjne*. The level of URL accessibility of was examined automatically using HTTP server responses.

Results and conclusions: Over 37% of examined papers contained Web citations. At the time of research over 70% of Web citations were still accessible, but this level was different for each journal. The half-life value of Web citations is approximately 7 years.

Originality/Value: The paper presents a comprehensive literature review on the problem of link rot which can be used as a background for further investigations related to this issue and as a reference for methodological considerations. According to the authors' knowledge, the issue of link rot had not been so far the subject of interest for Polish library and information science researchers. Therefore, the results of the research presented in the paper are the first attempt to describe the scale of this problem.

Keywords

Citation analysis. Polish scientific journals. Library and Information Science journals. The half-life value of Web citations. Web citations.

Dr MARCIN ROSZKOWSKI – adiunkt w Katedrze Informatologii na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Jego zainteresowania badawcze obejmują problematykę organizacji wiedzy w środowisku sieciowym, metadanych oraz technologii Sieci Semantycznej. Najważniejsze publikacje: M. Roszkowski: Kartoteka haseł wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF. W: Bibliografia – teoria, praktyka, dydaktyka, Warszawa 2016 (w druku); M. Roszkowski: Do czego potrzebne są nam ontologie? Charakterystyka funkcjonalna ontologii jako narzędzi reprezentacji wiedzy. W: Nauka o informacji w okresie zmian, Warszawa 2014, 151–167; M. Roszkowski: Od MARC 21 do Semantic Web: reprezentacja metadanych bibliograficznych w środowisku sieciowym. W: Bibliografi@: źródła – standardy – zasoby. Warszawa 2013, 5–25.

Kontakt z autorem:

m.roszkowski@uw.edu.pl

Katedra Informatologii

Wydział Dziennikarstwa, Informacji i Bibliologii

Uniwersytet Warszawski

ul. Nowy Świat 69

00–927 Warszawa

*Dr BARTŁOMIEJ WŁODARCZYK – adiunkt w Katedrze Bibliografii i Dokumentacji na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Jego zainteresowania badawcze dotyczą problematyki organizacji informacji i wiedzy. Najważniejsze publikacje: B. Włodarczyk: Mapy tematów jako podstawa semantyzacji języka haseł przedmiotowych. Warszawa 2016; B. Włodarczyk: Deskryptory formalne w Bibliotece Kongresu a tematy formalne w języku haseł przedmiotowych Biblioteki Narodowej i w języku KABA. *Przegląd Biblioteczny* 2014, 4, 518–536; B. Włodarczyk: *Topic Map as a Method for the Development of Subject Headings Vocabulary: An Introduction to the Project of the National Library of Poland. Cataloging & Classification Quarterly* 2013, 7, 816–829.*

Kontakt z autorem:

bm.wlodarczyk@uw.edu.pl

Katedra Bibliografii i Dokumentacji

Wydział Dziennikarstwa, Informacji i Bibliologii

Uniwersytet Warszawski

ul. Nowy Świat 69

00–927 Warszawa