Bartłomiej Siek

Biblio-Google-metria

Forum Bibliotek Medycznych 7/2 (14), 32-37

2014

Dr Bartłomiej Siek
Gdańsk – GUMed

BIBLIO-GOOGLE-METRIA

Abstract

The paper presents the changes in bibliometrics caused by the phenomenon of Google, including new search capabilities and data processing, as well as new types of documents setting up the flow of scientific information. The most common needs of bibliometrics data bases users' (representants of Google generation) are analyzed. Comparison of the results of the citation analysis of two Polish scholars representing different science areas with the examples of new "Google-possibilities" in scientific information processes not only demonstrates the need for further expansion of the definition of the document, but also a significant modification of understanding the phenomenon of citation.

Streszczenie

Referat stanowi przegląd zmian, jakie zaszły w bibliometrii wskutek pojawienia się fenomenu Google, obejmujących nowe możliwości wyszukiwania i przetwarzania danych, a także nowe typy dokumentów współtworzących obieg informacji naukowej. Analizie poddane zostały także najczęściej zgłaszane potrzeby informacyjne tych użytkowników bibliotecznego systemu informacyjnego, którzy stanowią już tzw. pokolenie Google. Porównanie wniosków wynikających z analizy cytowań dwóch polskich naukowców reprezentujących różne dziedziny z przykładami nowych „Google możliwości" w zakresie procesów informacji naukowej dowodzi konieczności nie tylko dalszego poszerzania definicji dokumentu, ale także znaczącej modyfikacji rozumienia fenomenu cytowania.

W tytule celowo nie użyto pojęć webometria[1] lub cybermetrics, bowiem celem artykułu jest określenie charakteru i zakresu zmian, jakie stały się udziałem pracowników bibliotek zaangażowanych w przygotowywanie dokumentacji bibliometrycznej. Mimo że formalnie aktem prawnym regulującym przygotowanie analizy bibliometrycznej jest rozporządzenie ministerialne wskazujące jako źródło danych bibliometrycznych bazę Web of Science, to akty prawne niższej rangi (wytyczne instytucji finansujących naukę, zarządzenia władz uczelni) oraz

[1] Najlepsze polskojęzyczne wprowadzenie do bibliometrii: Piotr Nowak: Bibliometria, webometria: podstawy, wybrane zastosowania. Wyd. 2 popr. Poznań: Wydawnictwo Naukowe UAM 2008 s. 138-159

uzus sprawiają, że bibliotekarze odpowiedzialni za sporządzanie dokumentacji bibliometrycznej zmuszeni są korzystać z innych jeszcze źródeł informacji.

Istnieje spora grupa zjawisk, które swe istnienie zawdzięczają Google'a, ale – niestety, w większości przypadków – nie stały się jednym z elementów praktyki bibliometrycznej. Przede wszystkim dotyczy to alternatywnych wskaźników bibliometrycznych, zwanych altmetrics, które próbują na nowo definiować miarę znaczenia w nauce. Warto jednak przypomnieć w tym miejscu, że za bardzo ważny krok w tym kierunku uznać należy wskaźniki obliczane na podstawie danych pochodzących z bazy Scopus (SNIP[2], SJR[3], na co odpowiedzią ze strony WoS jest inkorporowanie wskaźnika Eigenfactor[4]). Istniejące wskaźniki stosowane są do analizy nowych kanałów przepływu informacji (obliczanie indeksu H dla kanałów YouTube[5]), lub wiązane ze stosowanymi już narzędziami określania znaczenia w nauce (Twitter a przewidywanie cytowalności publikacji[6]).

Nie spełniły się pokładane w wyszukiwarkach Google'a nadzieje na to, że umożliwią one szybki rozwój mechanizmów otwartej nauki (tzw. zielona droga) za sprawą indeksowania repozytoriów tworzonych przez jednostki naukowe[7]. Przywołane poniżej przykłady mają zilustrować tezę, że w praktyce bibliometrycznej fenomen Google'a to kwestia jedynie skali. Więcej trzeba tłumaczyć, więcej trzeba gromadzić, mechanizm pozostaje ten sam. Już jedne z pierwszych analiz porównujących cytowania i sytowania (tak określa się przywołania w Internecie) wskazywały znaczącą korelację między

[2] Source Normalized Impact per Paper opracowany przez Centre for Science and Technology Studies (CWTS) na Uniwersytecie w Leiden, obliczany dla okresu trzech lat.

[3] SCImago Journal Rank wskaźnik opracowany przez badaczy z czterech hiszpańskich uniwersytetów: Granada, Extremadura, Carlos III (Madryd) and Alcalá de Henares, obliczany dla okresu trzech lat, waga przypisywana cytowaniom zależy od cytowalności czasopisma cytującego.

[4] Wskaźnik opracowany przez Jevina Westa i Carla Bergstroma z University of Washington, obliczany na podstawie danych z WoS dla okresu pięciu lat, waga przypisywana cytowaniom zależy od cytowalności czasopisma cytującego.

[5] Robert Hovden: Bibliometrics for Internet media: applying the h-index to YouTube. *J. Am. Soc. Inf. Sci. Technol.* 2013 T. 64 nr 11 s. 2326-2331

[6] Gunther Eysenbach: Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *J. Med. Internet Res.* 2011 T. 13 art. nr e123; Stefanie Haustein, Isabella Peters, Cassidy R. Sugimoto, Mike Thelwall, Vincent Lariviere: Tweeting biomedicine: an analysis of tweets and citations in the biomedical literature. *J. Assoc. Inf. Sci. Technol*. 2014 T. 65 nr 4 s. 656-669.

[7] Sytuację polskich repozytoriów w nawiązaniu do publikacji dotyczących Ameryki Łacińskiej analizuje Tomasz Lewandowski: Google Scholar a repozytoria i biblioteki cyfrowe w Polsce. [Dokument elektroniczny]. Dostępny w Internecie: http://otwartanauka.pl/analysis/case-studies/google-scholar-a-repozytoria-i-biblioteki-cyfrowe-w-polsce [dostęp: 02 IX 2014.]; tam też literatura przedmiotu.

wynikami analizy cytowań w bazie WoS i w zasobach sieci[8]. Tendencję tę potwierdziły kolejne analizy[9].

Otwartość, a właściwie nieograniczoność zasobów Internetu, za sprawą Google'a stała się elementem praktyki bibliometrycznej we wszystkich jej aspektach: udzielania informacji, gromadzenia danych, przygotowywania dokumentacji. Udzielanie informacji oznacza także weryfikację informacji, jakie do naukowców trafiają z redakcji czasopism proponujących publikację. Ponieważ coraz częściej za wydrukowanie artykułu autorzy płacą redakcji, umiejętne przygotowanie informacji o randze czasopisma mierzonej wskaźnikami bibliometrycznymi stało się strategią marketingową bardzo bliską manipulacji. Korzystając z faktu, że zastrzeżonym znakiem towarowym jest Journal Citation Reports, twórcy strony citefactor.org (nawet graficznie nawiązującej do produktów Thomson Reuters) stworzyli własny ranking czasopism na który powołują się redakcje czasopism. Zdarza się, że potencjalni autorzy zgłaszają się do biblioteki z prośbą o weryfikację danych; bywa, że do biblioteki trafia już kopia opublikowanej pracy z wydrukiem fragmentu listy z Journal Impact Factor. Obowiązek wyjaśnienia sytuacji spada na bibliotekarza.

Źródłem informacji o cytowaniach i indeksie H często przywoływanym przez naukowców, na których wniosek biblioteka przygotowuje analizy bibliometryczne, są internetowe platformy i portale społecznościowe[10], zliczające pobrania i odsłony publikacji umieszczanych przez autorów na swoich profilach. Ze względu na specyfikę wielu dziedzin dane dotyczące indeksu H wykazywane w tego rodzaju źródłach informacji bywają wyższe od danych w bazach WoS lub Scopus i zadaniem bibliotekarza staje się ponownie tłumaczenie specyfiki bibliometrii.

Po części wykorzystanie narzędzi pochodnych Google'a stało się obowiązkiem bibliotekarzy-bibliometrów za sprawą wytycznych dotyczących przygotowywania analizy bibliometrycznej na potrzeby konkursów grantowych organizowanych przez Narodowe Centrum Nauki. Załącznik nr 2 do Uchwały nr 52/2012 z dnia 14 czerwca 2012 r.[11], punkt D, podpunkt dotyczący paneli HS (nauki humanistyczne i społeczne) jako źródło informacji o sumarycznej liczbie cytowań i indeksie H wskazuje Web of Science lub Publish or Perish. Zapis taki wynika z faktu znaczącej przewagi ilościowej w bazie Web of Science czasopism reprezentujących dziedziny Science Technology Medicine (STM) nad czasopismami z zakresu humanistyki (Arts and Humanities – AH) i nauk społecznych (Social Sciences – SS). Co prawda uczelnie medyczne nie przeprowa-

[8] Liwen Vaughan, Deborah Shaw: Bibliographic and Web citations: what is the difference? *J. Am. Soc. Inf. Sci. Technol.* 2003 T. 64 nr 11 s. 1313-1322

[9] Mike Thelwall: Bibliometrics to webometrics. *J. Inf. Sci.* 2008 T. 34 nr 4 s. 605-621

[10] Np. academia.edu, researchgate.net, Google scholar

[11] http://ncn.gov.pl/userfiles/file/konkursy_ogloszone_2012-06-15/maestro-zal2.pdf [dostęp: 31 VIII 2014]

dzają postępowań awansowych w zakresie AH i SS, ale biblioteki tychże uczelni mają obowiązek zapewnić dostęp do informacji pracownikom także tych jednostek uczelni, które prowadzą badania właśnie w zakresie AH oraz SS. Publish or Perish to program stworzony przez Anne Harzing[12], który umożliwia przeprowadzenie analizy cytowań na podstawie zasobów wyszukiwarki Google Scholar[13]. O ile w przypadku analizy cytowań przeprowadzanej w WoS lub w bazie Scopus gwarantowany jest naukowy charakter źródeł cytujących, o tyle w przypadku analizy cytowań przeprowadzonej przy użyciu PoP wykaz źródeł cytujących może wymagać dodatkowej weryfikacji. Ta właśnie sytuacja podnoszona jest jako główny zarzut wobec stosowania PoP jako źródła informacji o wskaźnikach bibliometrycznych. Jednak porównanie analizy cytowań przeprowadzonej przy użyciu PoP i przy użyciu opcji Cited Reference Search w bazie WoS pozwala dostrzec te same problemy z uporządkowaniem danych uzyskanych jako wynik poszukiwań w obu źródłach informacji. Wykorzystanie zasobów Google Scholar jako podstawy analiz bibliometrycznych miałoby „ratunkiem" dla reprezentantów nauk humanistycznych i społecznych, dlatego jako przykład przeprowadzono analizę[14] cytowań publikacji Izabeli Koryś, socjologa, pracownika Instytutu Książki i Czytelnictwa Biblioteki Narodowej, która prowadzi badania dotyczące problemów migracji i problemów czytelnictwa. Wyniki przeprowadzonej analizy zawiera tabela 1. Bazy WoS i Scopus przeszukiwane przy użyciu opcji Author search nie notują żadnej pracy tej autorki; przeszukiwane przy użyciu opcji Cited reference serach (WoS) i References (Scopus) notują odpowiednio pięć pozycji cytowanych w sumie sześć razy (WoS) oraz dwadzieścia cztery pozycje cytujące (Scopus). Dalsze analizy uzyskanych wyników pozwoliły ustalić, że WoS Cited reference search de facto cytuje dwie prace, z których jedna notowana jest pod czterema wersjami tytułu, a Scopus w opcji Reference notuje dwadzieścia cztery prace cytujące, które przywołują dwadzieścia publikacji, dając w sumie trzydzieści trzy cytowania. Analiza przeprowadzona przy użyciu PoP daje następujący wynik: trzydzieści sześć dokumentów cytowanych dziewięćdziesiąt trzy razy, indeks H równy pięć. Dokładna analiza listy publikacji pozwala na połączenie różnych wersji cytowania tej samej pracy i okazuje się, że w grę wchodzi 27 publikacji.

Tak znacząca różnica między liczbą cytowań między bazami danych a Google Scholar może być oczywiście potraktowana jako wynik specyficznego statusu nauk humanistycznych i społecznych. Jednak także w przypadku reprezentantów nauk z zakresu STM różnice między wynikami analizy cytowań przeprowadzonej w WoS Author Search i WoS Cited Reference Search mogą być znaczące. Problemem nie okazuje

[12] Obecnie profesor w ESCP Europe (kampus w Londynie), wcześniej – kiedy stworzyła PoP – profesor University of Melbourne, por. http://www.harzing.com [dostęp: 31 VIII 2014]

[13] Obsługuje także Microsoft Academic Search

[14] Analizy cytowań przywoływane w artykule przeprowadzono w dniach 2 i 3 września 2014 r.

	liczba prac cytujących	liczba prac cytowanych	liczba cytowań	indeks H
WoS Author Search	0	0	0	0
WoS Cited Reference Search	6	2	6	1*
Scopus Author Search	0	0	0	0
Scopus Reference Search	24	20	33	3*
Publish or Perish	93	27	93	5**

Tabela 1: Analiza cytowań publikacji Izabeli Koryś

*obliczany na podstawie wyników wyszukiwań
**generowany automatycznie przez program

się wcale charakter dokumentów cytujących (jeden z częściej podnoszonych wobec PoP zarzutów), ale typ dokumentu cytowanego. Książki i rozdziały w książkach dopiero od niedawne są traktowane w bazach bibliometrycznych jako równorzędny wobec artykułu w czasopiśmie typ publikacji. W związku z powyższym naukowcy czynni już od kilkudziesięciu lat nie mają możliwości włączenia do analizy cytowań wyników uzyskanych przez ich publikacje w drukach zwartych, o ile nie skorzystają z opcje Cited Reference Search. W tabeli 2 przedstawiono wyniki analizy cytowań dla zapytania Roman Kaliszan w WoS Author Search i cytowań dwóch książek[15] tego naukowca zliczonych przy użyciu opcji Cited Reference Search.

WoS Author Search			WoS Cited Reference Search			suma	
liczba prac	cytowania	indeks H	liczba prac (całość)	liczba form przywołań 2 książek	cytowania 2 książek	cytowania	indeks H
244	5896	45	479	70	810	6706	47

Tabela 2: Analiza cytowań publikacji Romana Kaliszana

[15] Roman Kaliszan: Quantitative structure-chromatographic retention relationships. New York: John Wiley, 1987; Roman Kaliszan: Structure and retention in chromatography: a chemometric approach. Amsterdam: Harwood Academic Publishers 1997

Tej skali różnice w liczbie cytowań nie pozostawiają wątpliwości, że dążenie do dokładności wymaga porządkowania danych w przypadku każdego źródła informacji.

Analizy cytowań przeprowadzane w Google Scholar poddawane są krytyce i z tego jeszcze powodu, że PoP notuje nie tylko cytowania w artykułach dostępnych w sieci, ale także w innego rodzaju dokumentach, np. bibliograficznych bazach danych, bibliotekach cyfrowych itp. Przeprowadzona w WoS (Publication Name Search lub Cited Reference Search) analiza cytowań tytułu czasopisma wykazuje jednak, że drukowane w czasopismach bibliografie (czyli poprzedniczki bibliograficznych baz danych) traktowane są jako artykuły cytujące. Przykładowo, czasopismo Bibliotekarz cytowane jest w WoS Cited Reference Search czterdzieści cztery razy, w tym osiem cytowań w bibliografiach[16]; w analizie cytowań przeprowadzonej przy użyciu PoP Bibliotekarz uzyskał 252 cytowania, w tym jedno[17] w bibliograficznej bazie danych Conservation Information Network (BCIN). Skoro zatem wtórne dokumenty biblioteczne funkcjonują jako pełnoprawne źródło cytowań w WoS, to trudno podważać sensowność zliczania przez PoP notowania opisów publikacji w bibliograficznych bazach danych. Tym samym raz jeszcze należało by określić, co mieści się w ramach fenomenu cytowania (jest to po części powrót do krytyki IF i wyróżniania citable items w bazie WoS). Trafiające się coraz częściej w literaturze przedmiotu określanie Google Scholar mianem bazy danych należy uznać za impuls do uporządkowania kwestii terminologicznych już to na zasadzie puryzmu (wyszukiwarka nie może być bazą), już to redefinicji (przeszukiwalność oznacza istnienie zbioru danych).

Sama istota funkcjonowania wyszukiwarki Google jest de facto przeniesieniem do Internetu podstawowej idei bibliometrii (rozumianej jako badanie komunikacji międzydokumentowej), jaką jest notowanie wzajemnych odniesień. Algorytm wyszukiwawczy BackRub (nazwany następnie PageRank), uzupełnił stosowany przez inne wyszukiwarki mechanizm analizowania ilości wystąpień poszukiwanej frazy na stronie i uczynił z wyszukiwarki Google najpopularniejsze narzędzie wyszukiwawcze. Paradoksalnie, twórcy kierujący się przeciwnymi zasadami – WoS powołując się na prawo Bradforda z góry zakłada selekcję źródeł, Google dąży do uporządkowania światowych zasobów informacji – stawiają bibliotekarzy-bibliometrów przed tym samym problemem weryfikacji danych.

[16] Prace cytujące to: S. Richard Snodey: Information retrieval - comprehensive indexed bibliography of 1957-1961 world literature. *IEEE Transactions on Engineering Writing and Speech* 1964 T. EWS7 nr 1 s. 22nn; Bibliographia. *Rev. Int. Document.* 1965 T. 32 nr 2 s. 72nn

[17] Cytowana jest praca: E. Terlecki: Zaleszczotek książkowy. *Bibliotekarz* 1954 T. 21 s. 183-184