JAMES A. EVANS*

Publikacje elektroniczne a zawężenie obszaru poszukiwań nauki i wiedzy[1]

Dostęp do czasopism online niesie ze sobą obietnicę dostarczenia większej liczby informacji bardziej rozproszonym użytkownikom, pozwala także w sposób bardziej efektywny poszukiwać i przywoływać znaleziony materiał. Z uwagi na fakt, że wykorzystywany jest jednak odmiennie niż materiał drukowany – naukowcy i uczeni mają tendencje raczej do wykonywania poszukiwań elektronicznie i kierowania się odwołaniami do innych dokumentów powiązanych tematycznie (przez hiperłącza) niż do studiowania prac w wersji drukowanej – wykorzystanie czasopism elektronicznych może, paradoksalnie, zwiastować istotną zmianę w nauce. Wykorzystując bazę danych złożoną z 34 milionów artykułów, ich cytowań (od 1945 do 2005 roku) oraz badając ich dostępność online (1998-2005), wykazuję w niniejszej pracy, że w sytuacji, gdy coraz więcej wydań czasopism dostępnych jest w wersji elektronicznej, artykuły opatrzone odnośnikami do dzieł cytowanych obejmowały pozycje niedawne, z mniejszej liczby czasopism i artykułów, a większość z tych cytowań dotyczyła mniejszej liczby tytułów czasopism i artykułów. Dawniej dokładniejsze przeglądanie materiału drukowanego wymuszało wśród badaczy i naukowców pewne zakotwiczenie głęboko w przeszłości i współczesności nauki. Przeszukiwanie online jest wydajniejsze, a wykorzystanie hiperłączy szybko zapoznaje naukowców

* Począwszy od tego roku, redakcja pragnie w każdym roczniku zamieszczać przekład jakiegoś ważnego dla szeroko rozumianego bibliotekoznawstwa tekstu, który ukazał w naukowej prasie zagranicznej. Artykuł Jamesa A. Evansa otwiera nasz nowy dział.

[1] „Science" vol. 321, 18 lipca 2008.

z aktualnymi poglądami i opiniami, ale w rezultacie może przyspieszyć formowanie się naukowego konsensusu i zawęzić zasięg poszukiwań i prezentowanych myśli.

Odnosząc się do „bibliotek cyfrowych" i „technologii informacyjnej", nauka skupiła się na podkreśleniu korzyści i zalet, jakie oferuje elektroniczna możliwość wspomagania prac badawczych. Opublikowany ostatnio przez U.S. President's Information Technology Advisory Committee (PITAC, Komitet Doradczy Technologii Informacyjnych przy Prezydencie Stanów Zjednoczonych) raport panelowy zatytułowany *Biblioteki cyfrowe: uniwersalny dostęp do wiedzy* dobrze oddaje ten ton: „Wszyscy obywatele bez względu na to, gdzie i w jakim momencie się znajdują, mają prawo do używania wszelkich cyfrowych urządzeń posiadających połączenie z Internetem w celu zapewnienia sobie dostępu do wiedzy i zdobyczy ludzkości [...]. Zakreślona wizja przewiduje, że żadna izba klasowa, grupa ludzi czy jednostka nie będzie pozbawiona dostępu do zasobów światowej wiedzy i nauki"[2]. Takie stanowisko pomija specyficzny charakter interfejsu pomiędzy użytkownikiem a dostarczaną informacją[3]. Do tej pory nie podjęto szerokiej dyskusji na temat technologii przeglądania materiałów dostępnych online, metod i technik wyszukiwawczych czy, wreszcie, ich potencjalnego wpływu na naukę i wiedzę.

Współczesne badania bibliometryczne, zajmujące się praktyką wykorzystania materiału bibliotecznego, posługują się przy pomiarze wykorzystania zasobów drukowanych i elektronicznych badaniami ankietowymi, pomiarem logowań do baz danych, zapisem liczby wypożyczeń czy liczby pobrań z półek. Pomimo istniejących różnic w metodologii przeprowadzający takie badania zgadzają się co do jednego – wskaźnik wykorzystania materiału drukowanego ma tendencję spadkową, podczas gdy wykorzystania źródeł elektronicznych wzrasta[4] – przeciętny użytkownik woli pracować z materiałem dostępnym online niż tym w wersji drukowanej[5]. Przeprowadzone studia potwierdzają istnienie trzech najczęściej wykorzystywanych metod (praktyk) stosowanych przez autorów prac naukowych i badaczy publikujących swoje prace. Pierwsza ze stosowanych praktyk zakłada, że autorzy, powołując się na literaturę przedmiotu, przeglądają pobieżnie online lub

[2] R. Reddy et. al., *Digital Libraries: Universal Access to Human Knowledge (President's Information Technology Advisory Committee, Panel on Digital Libraries, 2001)*, www.nitrd.gov/pubs/pitac/pitac-dl-9feb01.pdf. Sprawozdanie to precyzuje wizję uniwersalnego dostępu, ale przyznaje jednocześnie, że warunkiem koniecznym jest udostępnienie „większej dostępności wysokojakościowej treści cyfrowej" przy zapewnieniu lepszej infrastruktury IT.

[3] M. McLuhan, *Understanding Media*, New York 1964, chap. 1.

[4] S. Black, „Libr. Resour. Tech. Serv." 49, 19 (2005).

[5] S.L. De Groote, J.L. Dorsch, „J. Med. Libr. Assoc." 91, 231 (2003).

kartkują pewną liczbę podstawowych tytułów ważnych dla danego przedmiotu, aby zorientować się w bieżącym stanie wiedzy na dany temat[6]. Gdy odpowiednie artykuły zostaną odnalezione w źródłach elektronicznych, są następnie często drukowane i studiowane w wersji papierowej[7]. Według drugiego sposobu najpierw dokonuje się wyszukiwań (tematycznych) w bazach danych dostępnych online. W ostatnich latach procentowy udział czasopism czytanych po uprzednim przeglądaniu materiału drukowanego zmalał i zastąpiony został analogicznymi wynikami poszukiwań online, szczególnie w przypadku badaczy najbardziej płodnych[8]. Wreszcie, aby przeglądać inne artykuły cytowane w pracach bądź z nimi związane, autorzy prac naukowych wykorzystują hiperłącza, które umieszczone są w elektronicznych wydaniach artykułów. Istnieją różnice pomiędzy dyscyplinami nauk, na przykład biolodzy wolą przeglądać materiał online, podczas gdy medycy wysoko cenią zapoznawanie się z materiałem drukowanym. Generalnie rzecz biorąc, naukowcy i uczeni studiują materiał drukowany, przeglądają drukowany i w wersji elektronicznej, a dokonują poszukiwań i zapoznają się z cytowaniami online[9]. Wynika to z dostępności czasopism drukowanych i elektronicznych. Zbiory czasopism drukowanych przechowywane są albo w jakimś fizycznym miejscu, posegregowane według tytułu czasopisma i tematu, według daty publikacji, albo w miejscu, gdzie znajdują się najnowsze publikacje. W przypadku czasopism drukowanych spis treści – lista tytułów artykułów i ich autorów – tworzy indeks podstawowy. Archiwa online z kolei pozwalają czytelnikowi szybko przeglądać odpowiedni materiał jednocześnie w wielu czasopismach, ułatwiają także przeszukiwania w całym zasobie dostępnych tytułów. W interfejsach użytkowników baz online, które oferują opcje przeszukiwania i dostępu do pełnego tekstu, takich jak na przykład 3 ProQuest, Ovid, EBSCO, JSTOR itp., opcja przeszukiwania zawsze umieszczona jest w widocznym miejscu w interfejsie, ponieważ liczba logowań wskazuje na jej częstsze użycie. Podczas dokonywania przeszukiwań ogólnego nietematycznego archiwum zawierającego referaty, tytuły, abstrakty, a czasami dokumenty pełnotekstowe, literatura do zadanego problemu może być wyszukiwana według zgodności z tematem i według daty. Z uwagi na to, że elektroniczne indeksowanie jest bogatsze,

[6] C. Tenopir, B. Hitchcock, S.A. Pillow, *Use and Users of Electronic Library. Źródło: An Overview and Analysis of Recent Research Studies*, Washington 2003.

[7] A. Friedlander, *Dimensions and Use of the Scholarly Information Environment: Introduction to a Data Set Assembled by the Digital Library Federation and Outsell*, Inc., Washington 2002, www.clir.org/pubs/reports/pub110/contents.html.

[8] P. Boyce, D.W. King, C. Montgomery, C. Tenopir, „Ser. Libr." 46, 121 (2004).

[9] C. Tenopir, D.W. King, A. Bush, „J. Med. Libr. Assoc." 92, 233 (2004).

autorzy, którzy nadal mogą przeglądać artykuły w ich wersji drukowanej, dokonują poszukiwań online[10].

Jaki jest efekt dostępności elektronicznych wydań czasopism online? Można by przypuszczać, że przez umożliwienie przeprowadzenia większej liczby przeszukiwań, poszukiwania dokonywane online mogłyby poszerzyć pracę opatrzoną cytowaniami w stosowne odnośniki do literatury i oddalić nieco badaczy, jako grupę kolektywną, od tzw. „core journals", tj. grupy tytułów czasopism podstawowych w danej dziedzinie w ich odpowiednich polach zainteresowań, a doprowadzić do rozrzuconych w różnych źródłach, ale indywidualnie relewantnych prac. Zamierzam w niniejszym artykule udowodnić, że nawet w sytuacji, gdy coraz starsze wydania archiwalne czasopism stają się dostępne online, naukowcy i uczeni cytują tylko artykuły, które ukazały się ostatnio – innymi słowy, choć ogólna liczba tytułów dostępnych online rośnie, cytowania związane są z coraz mniejszą ich liczbą.

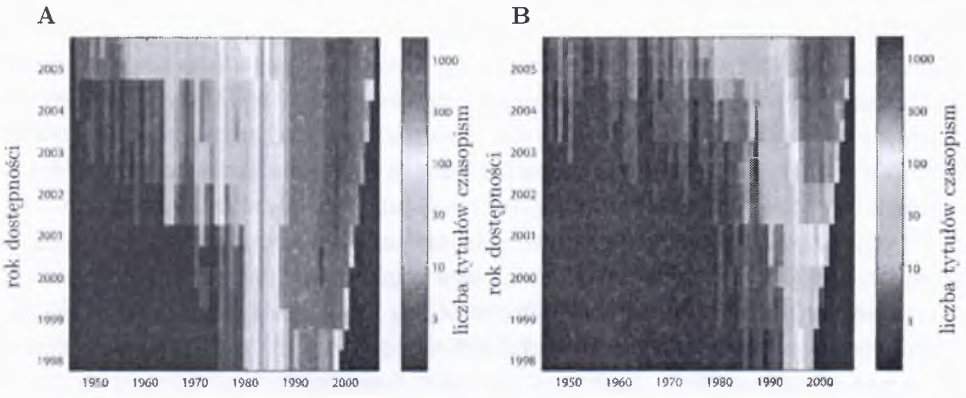
Dane dotyczące cytowań zaczerpnięte zostały z indeksów cytowań zamieszczonych w wydawanym przez Thompson Scientific indeksie cytowań *Science, Social Science and Arts and Humanities Citation Indexes*, najbardziej kompletnym źródle cytowań dostępnym na rynku. Dane z Citation Index (CI) aktualnie obejmują artykuły i powiązane z nimi cytowania z 6000 najczęściej cytowanych tytułów czasopism reprezentujących nauki ścisłe, nauki społeczne i humanistyczne od 1945 roku, dla ogólnej liczby ponad 50 milionów artykułów. CI oznakowuje (etykietuje) ponad 98% zamieszczonych tam tytułów czasopism, od jednego do trzech, specjalnymi kodami określającymi treść opisywanego artykułu, których jest 300. Etykiety te to, na przykład, „fizyka materii skondensowanej", „ornitologia" czy „chemia nieorganiczna i nuklearna". Schematy cytowań zostały następnie polinkowane z danymi śledzącymi dostępność czasopism online według *Fullest Sources Online* (FSO) sporządzanym przez Information Today, Inc.

FSO jest najstarszą i najobszerniejszą publikacją dotyczącą dostępności czasopism elektronicznych. Information Today rozpoczęło publikowanie FSO, ukazującego się dwa razy w roku, w 1998 roku. Publikacja obejmuje listę czasopism dostępnych w komercyjnie udostępnianych bazach danych (takich jak Lexis-Nexis, EBSCO czy Ovid) oraz tych, które dostępne są bezpłatnie na stronach internetowych wydawców lub czasopism. FSO informuje również, jakie lata obejmują numery archiwalne danego tytułu. W powiązaniu z danymi, których dostarcza numer ISSN (Międzynarodowy

[10] C. Shirky, *Ontology is Overrated: Categories, Links and Tags* (Clay Shirky's Writings About the Internet: Economics & Culture, Media & Community, Open Source, 2005), www.shirky.com/writings/ontology_overrated.html.

Znormalizowany Numer Wydawnictwa Ciągłego), dane z CI i FSO pozwoliły mi uchwycić, w jaki sposób dostępność online artykułu zmienia wykorzystanie opublikowanego materiału w kolejnych poszukiwaniach. Rozróżnienie źródła w FSO pozwala następnie na porównanie dostępu do materiału drukowanego z różnymi kanałami elektronicznymi, przez które autorzy prac naukowych docierają do artykułów – czy to prywatnie administrowany komercyjny portal czy też internetowy portal rozwijany w środowisku „open access". Połączone dane z CI-FSO dają razem liczbę 26 002 796 artykułów z czasopism dostępnych online do roku 2004 i odrębną grupę 8 090 813 pozycji (dodatkowo do liczby 26 milionów), które je cytowały i umieszczały w listach dzieł cytowanych. Rysunek 1 pokazuje tempo, z jakim nastąpiło przesunięcie w kierunku komercyjnych i bezpłatnych platform zapewniających dostęp do artykułów i w jaki sposób archiwizacja materiału uczyniła wcześniejszą literaturę przedmiotu dostępną w ostatnich latach.

Aby zbadać relacje pomiędzy dostępnością artykułów online a poziomem aktywności cytowań – przeciętnym poziomem głębokości cytowań, liczbą indywidualnych artykułów i tytułów czasopism, które były cytowane – oraz współczynnik koncentracji Herfindahla-Hirschmana cytowań w stosunku do indywidualnych artykułów i czasopism w zdefiniowanej przestrzeni czasowej, wykorzystano panelowe modele regresji (szczegóły zastosowanej metody w *Supporting Online Material* na internetowej stronie



Rysunek 1. Rozkład dostępności czasopism online, sporządzony na podstawie połączonych danych z ISI-FSO, poprzez **(A)** komercyjne subskrypcje oraz **(B)** bezpłatny dostęp ze strony internetowej czasopisma. Obszary „gorące" na wykresie odpowiadają wydaniom czasopisma publikowanym zaledwie kilka lat przed udostępnieniem ich online, np. w roku 2003 więcej czasopism było dostępnych (w bazach komercyjnych i bezpłatnie) – od roku 1999 odpowiednio około 1000 i 500 – niż z jakiegokolwiek innego roku. Wykres uwypukla tendencję wzrostową elektronicznych wydań czasopism od lat czterdziestych, pięćdziesiątych i sześćdziesiątych w latach 2004 i 2005

czasopisma „Science")*. Ponieważ badanie wykazuje znaczne różnice w odczycie i schematach badawczych stosowanych wobec danego obszaru, zastosowałem metodę statystyczną, zakładającą jednorodność analizowanych wyników, tak aby można było porównywać czasopisma i podpola wyłącznie wobec siebie w danym czasie, w sytuacji, gdy ich dostępność online zmieniała się. W ten sposób schemat cytowań w stosunku do danego tytułu czasopisma lub podpola był porównywany tylko, kiedy jego dostępność dotyczyła wydania drukowanego lub drukowanego i elektronicznego przez archiwum komercyjne (bazę danych) oraz online z dostępem bezpłatnym.

Pierwsze pytanie dotyczyło problemu, czy poziom głębokości cytowania – to jest lat między publikacją pracy a pracą, w której była cytowana – może być przewidywalny na podstawie zasięgu czasowego archiwalnych numerów online czasopisma, to jest ile lat wstecz obejmował dostęp do numerów archiwalnych dostępnych elektronicznie w ciągu poprzedniego roku, kiedy – jak można było założyć – autorzy przenosili i dołączali dzieła cytowane do swoich artykułów. W przypadku podpól obliczane to było jako lata od momentu pierwszej dostępności czasopisma. Dane zostały zebrane w tablicy publikacji, obejmującej okres 20 lat, a więc brano pod uwagę jedynie dane od roku 1965, tj. 20 lat od pierwszego roku obejmującego zbiór danych. Dla całego zbioru danych cytowania wskazywały na artykuły publikowane przeciętnie około 5-6 lat wcześniej (tabela S1). Przeciętna liczba lat dostępności online artykułów z czasopism to zaledwie 1,85 (dane obejmują lata od 1945), ale przy standardowym odchyleniu wynoszącym 5 lat i maksymalnym wyniku wynoszącym ponad 60 lat. Analizy dokonano przez rok cytowania i w ramach tytułu czasopisma oraz podpola. Przy generowaniu wszystkich wyników podlegających ewaluacji i wyznaczaniu linii trendu użyto metody najmniejszych kwadratów (tj. metody minimum sumy kwadratów błędów) dla statystycznej estymacji regresji linearnej.

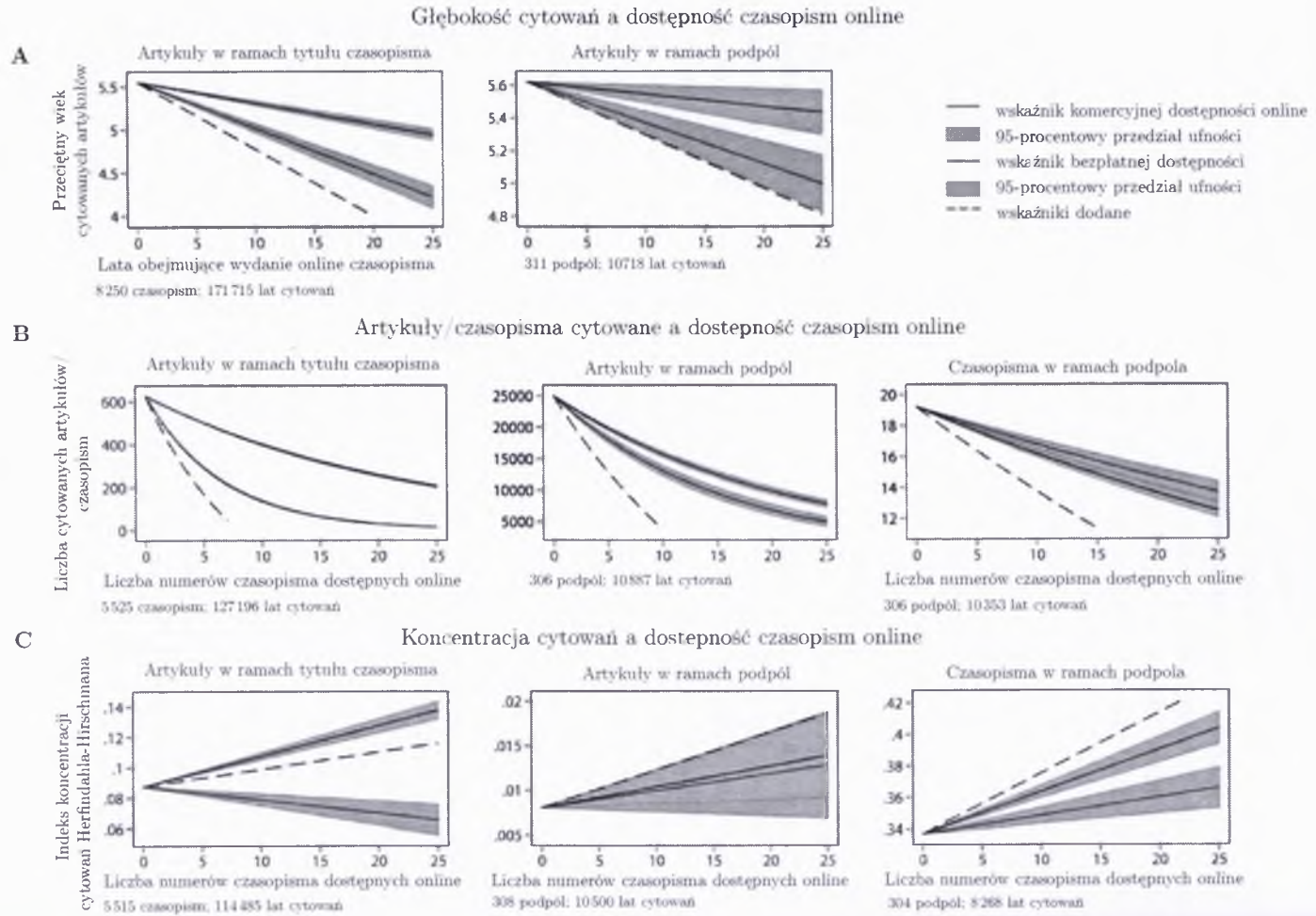
Wszystkie modele regresji zawierały zmienne wykorzystane następnie do statystycznej kontroli i uzasadnień, dlaczego cytowania mogą odnosić się do ostatnio publikowanych artykułów. Ciąg liczb całkowitych od 1 do 40, odpowiadający latom cytowań od 1965 do 2005 roku, wykorzystany został do wyjaśnienia ogólnej tendencji wzrostu liniowej funkcji trendu cytowań w czasie (szacunkowe dane dla tej zmiennej były zawsze dodatnie i statystycznie istotne, $P < 0{,}001$). W celu uzasadnienia przypuszczenia o dominującym udziale w cytowaniach stosunkowo niedawnego materiału w analizie uwzględniono zarówno średnią liczbę stron, jak i średnią liczbę odwołań do dzieł w cytowanych artykułach. O stosunkowo niedawnej publikacji świad-

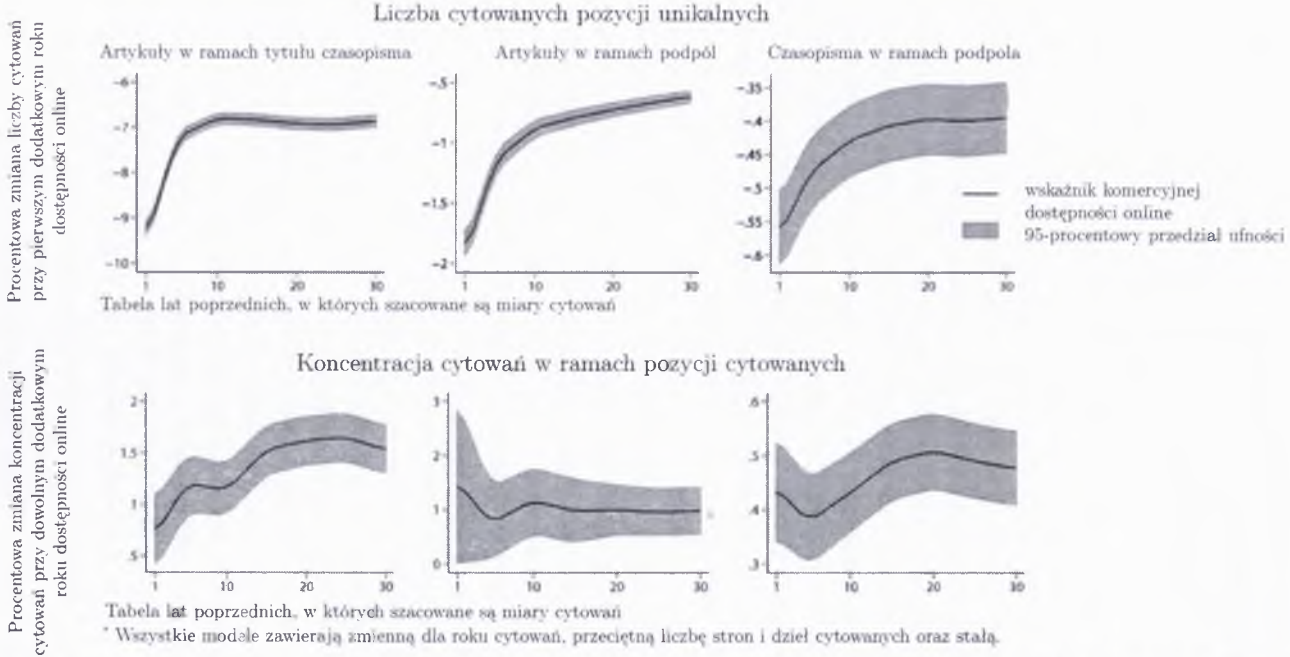
* Objaśnienie metody badań oraz tabele S1-S4 zob. http://www.sciencemag.org/cgi/content/full/321/5887/395/DC1.

czą długość artykułu (nowsze są krótsze) oraz liczba podanych źródeł. Przy czym liczba źródeł podawanych w najstarszych materiałach jest często ograniczana przez wydawców i w takich przypadkach estymacje dla stron były dodatnie, ale nie zawsze znaczące; podczas gdy te dla dzieł cytowanych były zawsze dodatnie i znaczne, $P < 0{,}001$: dłuższe artykuły z większą ilością podanej literatury odwoływały się do wcześniejszych prac. W modelach uwzględniono również wskaźnik średniego wieku słów tytułowych, aby rozważyć zakładaną możliwość, czy w ostatnich latach nauka i związane z nią badania nie skoncentrowały się na najnowszych (współczesnych) ideach lub zjawiskach odkrytych (bądź wynalezionych) niedawno. W tym celu wyznaczono wiek każdego słowa tytułowego w ramach odpowiedniej tablicy publikacji, np. poprzednich 20 lat, a następnie pomnożono go przez wartość wagową każdego słowa i w tytule j odpowiadającemu $\sum_{i=1}^{k}(1 + \ln(tf_{ij})) \times \ln\left(\frac{N}{df_i}\right)$, gdzie tf_{ij} równa się częstotliwości występowania terminu i w tytule j, a df_i równa się liczbie artykułów w danym roku, które zawierają termin i z ogólnej (całościowej) liczby artykułów danego roku N[11]. Takie podejście uprzywilejowuje wyróżniające pojęcia dystynktywne (słowa kluczowe) (np. fullereny, mikroRNA), natomiast znacznie niżej stawia terminy szersze znaczeniowo (np. gen, ocean) oraz praktycznie ignoruje tzw. stopwords, tj. słowa o uniwersalnym znaczeniu (np. and, the). Współczynniki regresji dla wskaźników wieku tytułów były zawsze dodatnie i znaczne, $P < 0{,}0001$. Wynika z tego, że tytuły zawierające starszą terminologię zamieszczały odnośniki do artykułów wcześniejszych. Każdy model regresji zawierał także stałą o znacznej ocenie ujemnej (spadek wartości zmiennej w czasie).

Wykresy przedstawione na rysunku 2 prezentują przebieg wpływu dostępu online, oszacowanego na podstawie całej próbki artykułów i zobrazowanego dla czasopism i podpól ze średnią liczbą cytowań, na liczbę cytowań. Rysunek 2A ukazuje jednoczesny efekt komercyjnej i bezpłatnej dostępności online na przeciętny wiek cytowań. Rozważmy przypadek czasopisma, którego artykuły podają w cytowanej literaturze wcześniejsze prace obejmujące przeciętnie 5, 6 lat wstecz – średnia próbki. Jeśli wydania tego czasopisma udostępnione zostają elektronicznie na dodatkowe 15 lat, zarówno w bazach komercyjnych, jak i w bazach bezpłatnych, przeciętny wiek dzieł cytowanych zmniejszy się do nieco poniżej 4,5 roku, ze spadkiem o 0,088 roku dla każdego nowego roku dostępności online. W ramach modeli podpól powtarzał się ten sam schemat, choć przedziały ufności były szersze (tabele S2-S4).

[11] C. Manning, H. Schütz, *Foundations of Natural Language Processing*, Cambridge 1999.


Głębokość cytowań a dostępność czasopism online
A
Artykuły w ramach tytułu czasopisma
Artykuły w ramach podpól
Przeciętny wiek cytowanych artykułów
Lata obejmujące wydanie online czasopisma
8 250 czasopism; 171 715 lat cytowań
311 podpól; 10 718 lat cytowań
wskaźnik komercyjnej dostępności online
95-procentowy przedział ufności
wskaźnik bezpłatnej dostępności
95-procentowy przedział ufności
wskaźniki dodane
B
Artykuły/czasopisma cytowane a dostępność czasopism online
Artykuły w ramach tytułu czasopisma
Artykuły w ramach podpól
Czasopisma w ramach podpola
Liczba cytowanych artykułów / czasopism
Liczba numerów czasopisma dostępnych online
5 525 czasopism; 127 196 lat cytowań
306 podpól; 10 887 lat cytowań
Liczba numerów czasopisma dostępnych online
306 podpól; 10 353 lat cytowań
C
Koncentracja cytowań a dostępność czasopism online
Artykuły w ramach tytułu czasopisma
Artykuły w ramach podpól
Czasopisma w ramach podpola
Indeks koncentracji cytowań Herfindahla-Hirschmana
Liczba numerów czasopisma dostępnych online
5 515 czasopism; 114 485 lat cytowań
Liczba numerów czasopisma dostępnych online
308 podpól; 10 500 lat cytowań
Liczba numerów czasopisma dostępnych online
304 podpól; 8 268 lat cytowań



Rysunek 2. Szacunkowy wpływ dostępności online artykułów (przedstawiony w latach dostępnych wydań online) na **(A)** średni wiek cytowań (opracowany za pomocą metody najmniejszych kwadratów i wyznaczonych współczynników regresji); **(B)** unikalną liczbę cytowanych czasopism (na podstawie ważonych wykładniczo współczynnikach ujemnego rozkładu dwumianowego maksymalnego prawdopodobieństwa); oraz **(C)** koncentrację Herfindahla-Hirschmana cytowań w ramach indywidualnych artykułów i czasopism (wyznaczonej na podstawie współczynników regresji metodą najmniejszych kwadratów). Każda z tych relacji zilustrowana jest względem średniej próbki wieku cytowań, liczby i koncentracji; każda przedstawiona relacja reprezentuje podstawowy model, który określa rok cytowania, liczbę stron oraz liczbę dzieł cytowanych w opatrzonych literaturą artykułach; podstawowy wiek cytowania model również określa średni ważony wiek ważonych słów w tytule w cytowanych artykułach. Szacunkowa zmiana procentowa, przy dodawanym jednym roku dostępności online, dla **(D)** liczby cytowanych unikalnych artykułów i tytułów czasopism oraz **(E)** koncentracji Herfindahla-Hirschmana w ramach tych cytowań, przy powiększeniu tablicy, w której ewaluowane są wskaźniki cytowań, od 1 do 30 lat, tj. od 1975 do 2005 roku.

W celu określenia wpływu dostępności online na wielkość cytowań unikalnych prac przebadałem zależność pomiędzy liczbą unikalnych artykułów a czasopismami cytowanych w danym roku cytowań względem głębokości dostępności online. Liczba indywidualnych artykułów i czasopism została obliczona na przestrzeni 20-letniego okresu, identycznie jak w poprzedniej analizie. Dla przeciętnego czasopisma liczba cytowanych artykułów wynosiła 632 artykuły rocznie, ale różnice były znaczne. Ponieważ wartości cytowań są dyskretne, a wysokie wartości koncentrują się wokół zbioru najważniejszych czasopism podstawowych, natomiast są bardzo zróżnicowane w przypadku pozostałych czasopism, relację tę dla dostępności online zmodelowano ujemnymi modelami dwumianowymi[12]. Model dwumianowy ujemny (ujemny rozkład dwumianowy – rozkład Pascala) jest uogólnieniem modelu Poissonowskiego i pozwala, w przypadku dodatkowego źródła wariancji ponad odpowiedni, na oczyszczenie błędu próbki. Jednorodność analizowanych danych dla tego modelu nie dotyczy oceny współczynnika kierunkowego liniowej funkcji trendu, lecz „parametru dyspersji", który wymusza na oszacowanej wariancji cytowań niezmienność w ramach czasopism czy podpól, pozwalając jednocześnie przyjmować jej każdą wartość w ramach tych grup. Modele zostały wyznaczone za pomocą metody maksymalnego prawdopodobieństwa i w ich wyniku powstały oceny współczynników regresji, które, gdy ważone wykładniczo, można zinterpretować jako stosunek (i) liczby unikalnych artykułów cytowanych po jednorocznym wzroście elektronicznego dostępu do czasopism do (ii) liczby artykułów cytowanych bez wzrostu dostępu online. Kiedy od tych relacji odejmiemy 1 i pomnożymy przez 100, to otrzymamy procentową zmianę jednorocznego wzrostu w dostępności online na liczbę odrębnych pozycji cytowanych. Wszystkie modele zawierały wskaźniki kontrolujące statystycznie rok, liczbę stron i liczbę dzieł cytowanych w artykułach zaopatrzonych w cytowaną literaturę przedmiotu.

W każdym kolejnym roku od 1965 do 2005 coraz więcej unikalnych artykułów było cytowanych z czasopism i podpól. Ogólna pula opublikowanego materiału naukowego cały czas rośnie i coraz więcej danych archiwizowanych jest corocznie w CI. Jednakże większa dostępność online nie potwierdza jednocześnie tej tendencji. Rysunek 2B ilustruje jednoczesny wpływ dostępności bezpłatnej i online na liczbę unikalnych (indywidualnych) artykułów cytowanych w czasopismach oraz liczbę unikalnych artykułów i czasopism cytowanych w podpolach. Panele przedstawiają te zależności dla przykładowego (hipotetycznego) czasopisma i podpola ze średnią próbki cytowań. Przy pięciu dodatkowych latach bezpłatnego i komercyjnego dostępu online liczba unikalnych cytowanych artykułów w ramach jednego

[12] J. Hausman, B.H. Hall, Z. Griliches, „Econometrica" 52, 909 (1984).

tytułu czasopisma spadłaby z 600 do 200; liczba artykułów cytowanych w ramach podpola zmalałaby z 25000 do 15000; a liczba czasopism cytowanych w ramach podpól zmalałaby z 19 do 16. Powyższe wyniki sugerują, że dostępność online może mieć wpływ ograniczający na liczbę unikalnych artykułów i czasopism cytowanych, doprowadzając go nawet do poziomu niższego od poziomu, który wystąpiłby, gdy czasopisma nie zostałyby udostępnione elektronicznie. Zapewnieniu bezpłatnego dostępu do wersji elektronicznej numerów czasopisma przez jeden dodatkowy rok towarzyszy jednocześnie 14-procentowy spadek cytowań unikalnych artykułów.

Mniej indywidualnych artykułów i czasopism doczekało się cytowań zaraz po tym, gdy zostały udostępnione w wersji elektronicznej. Choć wpłynęło to na ogólną koncentrację cytowań artykułów w nauce, to niecałkowicie i nie w pełni zdeterminowało proces. Cytowania może rozkładają się bardziej równomiernie na mniejszą liczbę artykułów, na które powołuje się szersze grono autorów publikacji naukowych. Aby oszacować stopień, w jakim dostęp online wpływa na koncentrację cytowań i ogranicza je do zaledwie niewielu artykułów (i czasopism), posłużyłem się indeksem koncentracji Herfindahla-Hirschmana, gdzie $\sum_{j=1}^{n}(s_j^2)$ oznacza procentowy udział cytowań s każdego artykułu j, podniesiony do kwadratu i zsumowany w ramach czasopisma lub podpola i w ciągu przebadanego okresu (20 lat). Koncentracja 1 wskazuje na to, że każde cytowanie do czasopisma i w danym roku jest do indywidualnego (unikalnego) artykułu; koncentracja nieco poniżej 1 sugeruje wysoki odsetek cytowań wskazujących na zaledwie kilka artykułów; a koncentracja zbliżająca się do zera sugeruje, że cytowania rozkładają się równomiernie, aktywizując dużą liczbę artykułów. W tej próbce koncentracje Herfindahla-Hirschmana artykułów cytowanych w czasopismach miały zasięg od 0,0000933 do 1, przy średniej 0,088 i dużym odchyleniu standardowym wynoszącym 0,195. Tam, gdzie artykuły nie były cytowane, nie można było obliczyć koncentracji. W celu sprawdzenia, czy koncentrację cytowań artykułów z ostatnich 20 lat można przypisać głębokości dostępności online, użyto modeli regresji liniowej. Podobnie jak poprzednio, modele zostały oszacowane dla artykułów w ramach tytułów czasopism oraz dla artykułów i czasopism w ramach podpól, przy założeniu dostępu zarówno do komercyjnych, jak i do niekomercyjnych elektronicznych baz danych. Koncentracje cytowań mają w przybliżeniu rozkład normalny (rozkład Gaussa), a użyte modele były poddane analizie za pomocą metody najmniejszych kwadratów.

Rysunek 2C przedstawia zbieżny w czasie wpływ dostępu do elektronicznych wydań czasopism, zarówno komercyjnych, jak i bezpłatnych, na koncentrację cytowań do indywidualnych artykułów i czasopism. Panel lewy,

z lewej strony wykresu, wskazuje, że liczba lat komercyjnej dostępności cytowań zdaje się znacząco zwiększać koncentrację cytowań do mniejszej liczby artykułów w ramach jednego czasopisma. Jeśli dodatkowe 10 lat wydawania czasopisma miałyby zostać umieszczone w dostępie online poprzez którekolwiek ze źródeł komercyjnych, model ten przewiduje, że jego koncentracja cytowań podniosłaby się z 0,088 do 0,105, a więc prawie o 20%. Bezpłatna dostępność do wydań elektronicznych miała nieco negatywny wpływ na koncentrację cytowanych artykułów w ramach grupy czasopism, ale marginalnie dodatni na koncentrację cytowanych artykułów w ramach podpól (panel środkowy) i zdawała się powodować znaczny wzrost koncentracji cytowań do zestawu najważniejszych czasopism w ramach podpól (panel prawy). Dostęp komercyjny miał stały pozytywny wpływ na koncentrację cytowań zarówno artykułów, jak i czasopism. Ogólne, zbiorcze podobieństwo pomiędzy dostępem komercyjnym i bezpłatnym dla wszystkich omawianych modeli sugeruje, że dostęp online – bez względu na charakter źródła – nadaje nowy kształt procesowi odkrywania wiedzy (ang. *knowledge discovery*) i jej wykorzystania w identyczny sposób. Dla wszystkich modeli otrzymano podobne wyniki, kiedy obecność czasopism w wieloskładnikowych archiwach komercyjnych (np. jednym, drugim, trzecim lub więcej) była brana pod uwagę i modelowana jednocześnie.

Chociaż okres 20 lat jest, zdawałoby się, wystarczającym okresem, aby rzetelnie przebadać wpływ dostępności online na cytowania, nie oddaje on jednak tendencji charakteryzującej ten wpływ. Na przykład, można wyobrazić sobie, że dostęp online powoduje wzrost unikalnej liczby artykułów cytowanych a zmniejsza koncentrację cytowań dla artykułów najnowszych, ułatwiając jednocześnie konwergencję do kanonicznych klasyków z bardziej odległej przeszłości. Aby prześledzić taką możliwość, przeprowadziłem identyczne analizy, ale obliczyłem zmienne z wzrastającego okresu sięgające od ostatniego roku obliczeniowego do ostatnich 30 lat. Aby próbki były porównywalne, dokonałem szacunkowej oceny wszystkich modeli obejmujących dane od roku 1975 (1945 plus 30-letni okres) do roku 2005, tak więc współczynniki 20-letniego okresu nie odpowiadają idealnie efektom zilustrowanym wcześniej. Szacunkowe procentowe zmiany, dotyczące liczby artykułów i czasopism cytowanych, i koncentracja cytowań wg indeksu koncentracji Herfindahla-Hirschmana, w ramach tych cytowań obliczone zostały jako skojarzone z jednorocznym przedłużeniem dostępności online. Powyższe estymacje oraz towarzyszące im 95-procentowe przedziały ufności przedstawione są na wykresach na rysunkach 2D i 2E. Zwiększony dostęp do archiwów online w roku poprzednim związany był ze spadkiem liczby cytowanych odrębnych artykułów w ramach czasopism oraz artykułów i czasopism w ramach podpól najbardziej w ciągu ostatnich lat

(rysunek 2D). Jednoroczna zmiana w dostępności online odpowiada 9-procentowemu spadkowi liczby artykułów cytowanych w ostatnim roku, ale tylko 7-procentowemu w cytowanych artykułach w ostatnich 20 i 30 latach. Schemat był identyczny dla artykułów i czasopism w ramach podpól (tabele S2-S4). Wpływ tabeli cytowań na koncentracje cytowań nie był już tak stały i konsekwentny (rysunek 2E). Niemniej jednak, w przypadku koncentracji artykułów w ramach podpól, wzrost indeksu koncentracji Herfindahla-Hirschmana był najwyższy – 1,5% rocznie przy dostępności online – i to obliczony dla literatury (dzieł cytowanych) wyłącznie do artykułów z ostatniego roku.

Zaprezentowane modele mają jednak wiele ograniczeń. Na przykład czasopisma takie jak „Science" wykorzystują Supporting Online Material for „Materiał and Methods" (Materiały pomocnicze do działu „Materiały i Metody" dostępne w Internecie), które często zawierają odnośniki do prac cytowanych, lecz nieindeksowanych przez CI. Teoretycznie jest zatem możliwe, choć mało prawdopodobne, że te odwołania do literatury dotyczą wcześniejszych lub bardziej różnorodnych artykułów. Co więcej, poddając badaniu jedynie standardowe tytuły czasopism, badania nasze nie są w stanie oddać pełnego obrazu współczesnych mediów oddanych do dyspozycji badaczom, takich jak naukowe blogi, serwisy oparte na mechanizmie wiki czy wreszcie inne internetowe serwisy korzystające z modeli alternatywnych form recenzji naukowej. Te nowe media prawie bez wątpienia linkują do absolutnie najnowszych osiągnięć w nauce – często właśnie przez linki internetowe[13] – ale mogą również wskazywać na bardziej zróżnicowany źródłowo materiał.

Ogólnie rzecz biorąc, przedstawione modele wykazują, że gdy archiwalne numery czasopisma zostają udostępnione online, albo przez dostawcę komercyjnego, albo z dostępem bezpłatnym, następuje zmiana we wzorze (modelu) cytowań. Gdy coraz starsze archiwalne numery zostają udostępnione w trybie online, dzieła cytowane obejmują nowsze artykuły; gdy coraz więcej artykułów udostępnionych zostaje online, mniej tych artykułów jest cytowanych, a cytowania stają się bardziej skoncentrowane, obejmując mniejszą liczbę artykułów. Zmiany te najprawdopodobniej oznaczają, że przesunięcie punktu ciężkości z przeglądania materiału drukowanego na poszukiwania dokonywane online sprzyja pomijaniu starszej i mniej relewantnej literatury przedmiotu. Co więcej, hiperłącza umieszczone w archiwach dostępnych online kontaktują autorów publikacji naukowych z naukowym konsensusem co do tego, co jest najistotniejsze w pracach wcześniejszych – która praca jest szeroko dyskutowana i często

[13] R. P. Dellavalle et al., „Science" 302, 787 (2003).

cytowana. Przy obu zastosowanych strategiach naukowcy korzystający z materiału online omijają wiele z artykułów mniej związanych z tematem, do których ci dokonujący przeszukiwań na materiale drukowanym jednak znajdują dojście. Jeśli dokonujący przeszukiwań online łatwiej mogą dotrzeć do aktualnie obowiązujących i przeważających opinii, jest też bardziej prawdopodobne że zaakceptują je i, podążając wyznaczoną już drogą, sami umieszczać będą w swoich pracach cytowania odnoszące się do mniejszej liczby artykułów. Badania nad skrajnym nierównouprawnieniem internetowych hiperłącz[14], naukowych cytowań[15] oraz innych form mechanizmu przyciągania preferencyjnego (ang. *preferrential attachment*)[16] sugerują, że niejednorodne różnice jakościowe ulegają jeszcze dodatkowemu wzmocnieniu, kiedy agenci reprezentujący platformy cyfrowe udostępniające bazy danych stają się świadomi tych wyborów. Agenci przyjmują wybór profesjonalistów jako informację, która w konsekwencji determinuje ich wybór – jako znak jakości – i wkalkulowują je w swoje własne selekcje literatury i cytowań. Umożliwiając naukowcom szybkie dotarcie i konwergencję z przeważającą opinią, czasopisma elektroniczne przyspieszają naukowy konsensus. Ale pośpiech może kosztować więcej niż subskrypcja elektronicznego archiwum online: odkrycia i idee, które nie znajdują szybkiego i szerokiego poparcia i naukowego konsensusu, szybko też bywają zapomniane.

Przedstawione wyniki badań wyraźnie sugerują, że, paradoksalnie, jedną z głównych wartości poszukiwań opartych na materiale drukowanym jest ich słabe indeksowanie. Słabe indeksowanie (i pozycjonowanie), tj. indeksowanie jedynie przez tytuł i nazwisko autora, przede wszystkim w obrębie zestawu czasopism podstawowych, prawdopodobnie miało, niezamierzone zresztą, konsekwencje, które wspomagały integrację nauki i wiedzy. Przeprowadzanie naukowców przez niebezpośrednio powiązane artykuły, przeglądanie i studiowanie materiału mogło ułatwiać szersze porównanie i pozwalało zapoznać się z wcześniejszymi ideami. Współczesne praktyki obowiązujące na wyższych uczelniach znajdują paralelę w zaistniałym przesunięciu w publikacjach – krótszych w latach odwołań, bardziej wyspecjalizowanych w swoim zakresie, doprowadzających ostatecznie do powstania już mniej prawdziwych dysertacji, lecz raczej albumu artykułów[17].

Przejście do wiedzy i nauki opartych na źródłach internetowych, wydaje się odzwierciedlać jeszcze jeden krok na drodze zainicjowanej o wiele wcześniej-

[14] A.L. Barabási, R. Albert, „Science" 286, 509 (1999).

[15] R.K. Merton, „Science" 159, 56 (1968); D.J. de Solla Price, „Science" 149, 510 (1965).

[16] H.A. Simon, „Biometrika" 42, 425 (1955); M.J. Salganik, P.S. Dodds, D.J. Watts, „Science" 311, 854 (2006).

[17] J. Berger, *Exploring ways to shorten the ascent to a Ph.D.*, „New York Times", 3 October 2007; www.nytimes.com/2007/10/03/education/03education.html.

szym przejściem z monografii ujętych w szerszym kontekście tematycznym, takich jak *Philosophiae naturalis principia mathematica* Newtona[18] czy *O pochodzeniu gatunków* Darwina[19], na współczesny artykuł naukowy. Przywołane dzieła powstawały w okresie przekraczającym dziesięciolecie, nie tylko były mocno zaangażowane w naukowe debaty swego okresu, ale ich autorzy również wprowadzali swoje propozycje w dyskurs z astronomami, geometrami czy przyrodnikami z wieków poprzednich. Naukowcy i uczeni XXI wieku wykorzystują przeszukiwania online oraz hiperlinkowanie, aby formułować i publikować swoje argumentacje w sposób bardziej efektywny, wplatają je w bardziej zogniskowane – ale i tym samym bardziej zawężone – przeszłość i teraźniejszość.

Chciałbym wyrazić wdzięczność za pomoc w badaniach uzyskaną z grantu nr 0242971 NSF, dane z Science Citation Index uzyskane od Thompson Scientific, Inc. oraz dane z *Fulltext Sources Online* od Information Today, Inc. Chciałbym również podziękować p. J. Reimer za pomocne uwagi i spostrzeżenia.

Przeł. Tomasz Olszewski

[18] I. Newton, *Principia*, wyd. 4, New York 1883 (pierwszy raz opublikowane w 1687).
[19] C. Darwin, *The Origin of Species*, New York 1867 (data pierwszej publikacji – 1859).