PRZEGLĄD BIBLIOTECZNY 2011 z. 4
PL ISSN 0033-202X

WŁADYSŁAW MAREK KOLASA
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytet Pedagogiczny w Krakowie
e-mail: wmkolasa@gmail.com

RETROSPEKTYWNY INDEKS CYTOWAŃ W HUMANISTYCE. KONCEPCJA, METODA, ZASTOSOWANIA

Władysław Marek Kolasa jest adiunktem w Instytucie Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Pedagogicznego w Krakowie. Naukowo zajmuje się problematyką historii i teorii mediów oraz naukometrią. Od 2003 r. jest członkiem zarządu Komisji Prasoznawczej Polskiej Akademii Nauk oraz zasiada w redakcji „Rocznika Historii Prasy Polskiej". Zajmuje się ponadto systemami informatycznymi i bibliotekarstwem cyfrowym; pracował m.in. nad wdrożeniem serwisów: *Fidkar Małopolski* oraz *Małopolska Biblioteka Cyfrowa*. Opublikował ponad 70 prac naukowych w tym 4 książki. Do ważniejszych należą: *Prasa Krakowa w dekadzie przemian* (Kraków 2004), *Polska bibliografia prasoznawcza 1996-2001* (Kraków 2005) i *Katalog małopolskich mediów lokalnych i regionalnych* (Kraków 2005).

SŁOWA KLUCZOWE: Indeks cytowań. Bazy danych. Projektowanie. Metodologia. Humanistyka

ABSTRAKT: **Teza/cel artykułu** – W artykule opisano metodę budowy indeksu cytowań, uzyskanego poprzez przekształcenie bibliograficznej bazy danych. Przedstawiono poszczególne etapy tworzenia takiej bazy i poddano je krytycznej ocenie. Opisano kolejno fazy: projektowania systemu, metodykę tworzenia bazy danych, tworzenia cytowań ich konwersji oraz eksploatacji. **Metody badań** – Tekst opiera się na autorskich doświadczeniach zebranych podczas tworzenia Indeksu Cytowań Historiografii Mediów Polskich (ICHMP). W opracowaniu wykorzystano wskaźniki bibliometryczne, zaczerpnięte z ICHMP oraz (kontekstowo) z literatury przedmiotu. **Wyniki** – Zbudowany według tej metody indeks (ICHMP) zawierał 24 627 dokumentów powiązanych siatką 63 811 cytowań. Największą zaletą opisanej koncepcji jest wysoka skuteczność. W bazie osiągnięto wskaźniki porównywalne do indeksów filadelfijskich, np. citation impact' wyniósł 6,7; maksimum cytowań na 1 publikację – 415, zaś autor-rekordzista zgromadził 1075 cytowań. **Wnioski** – Przedstawiona koncepcja (umownie: indeks retrospektywny) może być z powodzeniem zastosowana w innych dziedzinach humanistycznych. Głównym walorem rozwiązania są: niskie koszty, duża skuteczność oraz ogromne możliwości obliczeniowe.

WSTĘP

Korzyści wynikające z wykorzystania indeksów cytowań są niewątpliwe. Można je użytkować w wielu obszarach i funkcjach, jak też rozmaitych zastosowaniach (praktycznych i strategicznych). Najczęściej jednak wyodrębnia się cztery podstawowe pola ich eksploatacji: (1) jako inteligentne narzędzie

wspomagające tradycyjne wyszukiwanie; (2) jako mechanizm ewaluacji prac naukowych (generowania rangi publikacji, autora, zespołu); (3) jako bazę do tworzenia schematów powiązań frontów badawczych, czyli źródło do kartografowania nauki oraz (4) jako metodę generowania powiązań prac podobnych przy użyciu współcytowań (ang. *co-citation, bibliographic coupling*). W praktyce można wskazać wiele węższych zastosowań (np. ocena czasopism na potrzeby gromadzenia), a także zastosowań szerszych, jak np. wykorzystanie cytowań do wizualizacji rozwoju określonej dyscypliny naukowej. Spektrum możliwych rozwiązań jest doprawdy ogromne, o czym przekonuje np. lektura artykułów opublikowanych w *Scientometrics*, *Journal of Informetrics* czy *JASIST*. Idea, która legła u podstaw budowy indeksu cytowań, czyli kognitywny związek pracy cytującej z pracą cytowaną (Garfield, 1955) znalazła także szerokie zastosowanie w systemach informatycznych, szczególnie przy projektowaniu serwisów do przetwarzania i ewaluacji informacji internetowej; przykładami są *CiteSeerX* i *Google Scholar* (Lawrence et al., 1999; Bakkalbasi et al., 2006).

PROBLEMY REPREZENTATYWNOŚCI DANYCH O CYTOWANIU LITERATURY NAUKOWEJ

Możliwości, jak przekonuje praktyka, są ogromne, lecz nie można przemilczeć także kilku ograniczeń w tym zakresie. Problemy sprowadzają się w zasadzie do dwu głównych kwestii: (1) dostępności danych wysokiej jakości oraz (2) możliwości systemów, które je przetwarzają. O ile ostatni problem ma charakter czysto techniczny i sprowadza się do zwiększenia funkcjonalności interfejsu, o czym przekonuje ewolucja platformy *Web of Knowledge*, o tyle kwestia jakości danych jest wciąż dużym, nierozwiązanym problemem. Skalę problemu ujawnia analiza zawartości głównych światowych indeksów (baz *Thomson ISI* oraz indeksu *Scopus*). W obu przypadkach występuje tutaj rażąca nadreprezentacja nauk przyrodniczych, które stanowią ok. 90% zawartości baz, podczas gdy na nauki społeczne przypada – niespełna 8%, a humanistyczne – tylko 2,5% (Marshakova-Shaikevich, 2009, s. 194). Stan taki odbija się niekorzystnie także na wskaźnikach, np. *citation impact'* (tj. liczbie cytowań na jedną publikację cytowaną), który dla korpusu *science & technology* oscyluje w granicach 5,02-18,59, podczas gdy dla *social science* jest średnio dwukrotnie niższy 2,80-6,48, a w przypadku *humanities* ledwie symboliczny 1,35-1,80. W takim kontekście łatwo o konkluzję, że jedynie w przypadku nauk ścisłych, medycznych i technicznych narzędzia te prowadzą do uzyskania względnie poprawnych wyników. Wskazany problem pogłębia także silna polaryzacja językowa publikacji i niereprezentatywność geograficzna indeksowanych źródeł. Ostatnią tezę ilustruje tab. 1, w której zestawiono wskaźniki reprezentacji czasopism w bazach *ISI* w porównaniu z *Ulrich's International Periodical Directory*. Z zestawienia łatwo odczytać przygniatającą przewagę USA i Wielkiej Brytanii nad pozostałymi państwami (razem 59% w zakresie nauk ścisłych i technicznych (NST) oraz aż 77% w zakresie nauk społecznych i humanistycznych NSH), nadto silną nadreprezentację pism z tych krajów w obu korpusach tytułów (19-36% NPT i 35-55% NSH).

Tabela 1

Czasopisma naukowe w serwisach *Thomson ISI* [liczba tytułów]
(Archambault & Gagné, 2004, p. 19)

Państwo	Nauki ścisłe i techniczne [NST]			Nauki społeczne i humanistyczne [NSH]		
	Thomson ISI (%)	Ulrich (%)	Różnica (%)	Thomson ISI (%)	Ulrich (%)	Różnica (%)
United States	36	31	19	50	37	35
United Kingdom	23	17	36	27	18	55
Netherlands	9,4	8,3	14	7,7	7,4	5
Germany	7,7	6,2	25	3,9	5,9	-34
Switzerland	2,7	2,1	26	0,6	0,5	8
France	2,4	2,6	-6	1,0	1,4	-24
Japan	2,3	3,7	-39	0,5	1,0	-55
Russian Federation	1,6	1,4	12	0,3	0,3	36
Canada	1,3	1,3	1	2,5	3,2	-21
Australia	1,2	2,1	-42	1,1	3,6	-71
Italy	1,1	1,7	-38	0,1	1,2	-89
Spain	0,4	1,3	-72	0,3	1,0	-75
India	0,9	2,2	-61	0,2	1,6	-86
China	0,9	2,9	-69	0,1	0,9	-91
Belgium	0,2	0,4	-52	0,5	2,1	-75
Poland	0,7	1,6	-58	0,2	1,3	-87
Brazil	0,3	1,1	-72	0,04	1,0	-96
Other	7,5	14	-45	3,5	13	-73

W piśmiennictwie naukowym od wielu lat trwa dyskusja na temat wskazanych dysproporcji i przywoływane są rozmaite argumenty (Nowak, 2008, s. 27). Najczęściej podkreśla się fakt, że u podłoża takiego stanu rzeczy tkwi pojęciowe i metodologiczne ukształtowanie poszczególnych nauk, czyli ich paradygmat (Nowak, 2004). Istnieją zatem dyscypliny, które ze swej istoty mają charakter międzynarodowy (przyrodnicze i formalne), czyli nauki najbardziej abstrakcyjnie zredukowane[1]; inne wykazują mniejszą skłonność do funkcjonowania w obiegu międzynarodowym, gdyż są bardziej konkretne i odwołują się do lokalnego środowiska (np. nauki społeczne i niektóre stosowane); wreszcie – nauki będące z natury lokalne, których przedmiot badań jest niejako zespolony ze środowiskiem, w jakim powstaje – czyli nauki humanistyczne. Dodajmy, że związek ten w humanistyce jest na tyle silny, że w przypadku większości dyscyplin piśmiennictwo naukowe powstaje niemal w całości w językach narodowych (Kolasa, 2011b), np. dla historii udział języka narodowego waha się w przedziale 95,6-98,8% (Francja – 98,8; USA – 98,8; W. Brytania – 98,7;

[1] Na problem ten zwracają uwagę epistemolodzy (np. A. Huxley). Zwykle podkreślane są dwa aspekty tej redukcji: przedmiotowy i metodologiczny. W szczególności: (1) nauki realne są tematycznie zredukowane, gdyż ograniczają swój temat (przedmiot) do określonego aspektu oraz (2) nauki realne są metodycznie abstrakcyjne, ponieważ wycinek, którym się zajmują, o tyle tylko wchodzi w ich zasięg, o ile pozwala na to ich określona metoda – szerzej Anzenbacher, 1992, s. 25-29.

Włochy – 98,7; Niemcy – 96,1; Polska – 96,0; Hiszpania – 95,8; Rosja – 95,6). Uzasadnienie tej tezy nie jest szczególnie trudne. Zdecydowana większość nauk historycznych i filologicznych wymaga zarówno od autora, jak i czytelnika szerokiego spektrum wiedzy kontekstowej (kultura, polityka, język i in.), niemożliwej do przyswojenia na drodze innej niż czynne uczestnictwo w danej kulturze i trudnej do wyrażenia w języku innym niż narodowy. Humanistykę cechuje też szereg innych osobliwości, które stawiają ją w opozycji do nauk przyrodniczych. Cechą taką jest np. wysoka pozycja książki, jako głównej formy prezentacji wyników oraz wysoki udział prac jednoautorskich.

W tym kontekście nie dziwi, że na świecie podjęto liczne próby tworzenia lokalnych indeksów cytowań. Na różnym poziomie zaawansowania znajdują się aktualnie m.in. projekty *Spanish Social Sciences Journals* (Torres-Salinas et al., 2009); *The Taiwan Humanities Citation Index* (Chen, 2004); *African Citation Index Project* (Nwagwu, 2010); *Indian Citation Index*[2] (Giri & Das, 2011) czy *European Citation Index for the Humanities* (Di Donato, 2004). Do udanych i jednocześnie największych przedsięwzięć tego typu należy rozwijany od 2000 r. *Chinese Social Sciences Citation Index*[3], który obejmuje prace z 2700 czasopism i jest z powodzeniem wykorzystywany do oceny nauki przez rząd chiński (Hua, 2001; Gong et al., 2007).

Od połowy lat 90. XX w. eksperymentalne projekty powstały też w Polsce; zbudowano m.in.: *Indeks Cytowań Socjologii Polskiej* (Wincławska & Wincławski, 1995), indeks *ARTON* (Waga & Drabek, 2002) oraz indeks CYTBIN (Stefaniak & Swoboda, 2004; Nowak, 2008, s. 46-53); w literaturze rozpoczęła się też dyskusja na ten temat (Webster, 2001; Nowak, 2000, 2001, 2004). Największe nadzieje wiązano z tworzonym od 1998 r. indeksem *Polska Literatura Humanistyczna ARTON*; niestety projekt po początkowym okresie wzrostu wszedł w stan stagnacji; w 2004 r. liczył 7481 dokumentów i 124 967 cytowań; w 2009 – ok. 11 500 i 173 tys. (Drabek & Waga, 2009), a dwa lata później wzrósł o niespełna 250 dokumentów i 1000 cytowań. Wskazany proces nie jest jednak zamknięty, wciąż powstają projekty nowe, czego przykładem jest inicjatywa przebudowy serwisu *BazTech* w indeks cytowań (Derfert-Wolf et al., 2005) oraz serwis *BazEkon* dynamicznie rozwijany na Uniwersytecie Ekonomicznym w Krakowie (Osiewalska, 2008). Wspólnym rysem wyliczonych projektów jest ich wycinkowy charakter (najczęściej ograniczony niewielką liczbą ujętych tytułów lub wąskim zasięgiem chronologicznym), co ostatecznie prowadzi do niskich wskaźników (np. *citation impact*). W efekcie indeksy polskie najczęściej dostarczały materiału do prostych analiz bibliometrycznych (np. Drabek & Tomaszczyk, 2008) lub studiów komparatystycznych (np. Webster, 2001). Nie powstała natomiast na ich podstawie żadna praca operująca wskaźnikami porównywalnymi do indeksów filadelfijskich. W świetle tego brak było argumentów, aby sfalsyfikować tezę, że niskie wskaźniki dla humanistyki w SCI (SCI/SSCI/A&HCI) są jedynie wynikiem niskiej reprezentacji obszaru *humanities* w tych źródłach.

W 2005 r. w Instytucie Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Pedagogicznego w Krakowie w zespole kierowanym przez autora niniejszego artykułu podjęto pracę nad bazą *Indeks Cytowań*

[2] *Indian Citation Index*: http://www.indiancitationindex.com/ [2011.06.11].
[3] *Chinese Social Sciences Citation Index*: http://www.cssci.com.cn/eindex.htm [11.06.2011].

Historiografii Mediów Polskich [ICHMP][4], w którym rozpoczęto rejestrację w miarę kompletnego zestawu literatury naukowej na tytułowy temat za lata 1945-2009[5]. Ponieważ nie udało się uzyskać zewnętrznych funduszy na finansowanie projektu, całość prac oparto na rozwiązaniach, które nie wymagały angażowania większych środków finansowych (wykorzystano dostępną infrastrukturę IINiB UP oraz przychylność firmy Sokrates Software, która zapewniła dostęp do serwera z bazą danych). Na potrzeby projektu zastosowano ponadto eksperyment polegający na przekształceniu bazy bibliograficznej w indeks cytowań. Po pięciu latach prac, w których okresowo uczestniczyli przeszkoleni studenci najwyższych lat studiów bibliotekoznawczych (razem ok. 500 osób) bazę ukończono. W wersji finalnej baza liczyła 24 627 dokumentów powiązanych siatką 63 811 cytowań, a główny trzon stanowiły opisy dotyczące historii mediów (15 920 dokumentów cytowanych 52 254 razy, w tym 46 152 bez autocytowań). Efekty prac przeszły najśmielsze oczekiwania twórców, gdyż uzyskano wskaźniki zbliżone do SCI (tab. 2). Pozwoliło to już na wstępie zakwestionować tezę, jakoby humanistyka z natury swej cechowała się niskimi wskaźnikami cytowalności; uzyskano bowiem empiryczny dowód, że humanistyka (na przykładzie historii) zachowuje się w sposób zbliżony do niektórych nauk przyrodniczych (np. materiałoznawstwa), nie zaś tak, jak opisują ją dane A&HCI.

Tabela 2
Indeks Cytowań Historiografii Mediów Polskich a Science Citation Index[6] (Kolasa, 2011a)

Wybrane wskaźniki	ICHMP		SCI / A&HCI					
	Historia		Całość		Materiało-znawstwo		Historia	
	N	%	N	%	N	%	N	%
Liczba pozycji	15041		5655186		191128		b.d.	
Liczba cytowań (bez autocytowań)	46152		40516820		575725		b.d.	
Liczba poz. cytowanych	6924	46,0	3768822	66,6	97807	51,1	b.d.	19,6
Liczba poz. niecytowanych	8117	53,9	1886364	33,3	93321	48,8	b.d.	80,4
Śr. liczba cytowań 1 poz.	3,06		7,16		3,01		0,37	
Śr. liczba cytowań 1 poz. cytowanej	6,66		10,75		5,89		1,91	

Na kolejnych etapach pracy nad bazą ICHMP wyposażono ją w niezbędne mechanizmy obliczeniowe, które dostarczyły materiału empirycznego, umożliwiającego weryfikację praw bibliometrycznych oraz wizualizację trendów w rozwoju badanej dyscypliny (Kolasa, 2011a).

Rozwiązanie, które zastosowano do budowy ICHMP (nazwane umownie retrospektywnym indeksem cytowań) może być – zdaniem autora – z powodzeniem zastosowane w innych dziedzinach humanistycznych.

[4] Bardzo okrojoną wersję bazy z interfejsem w języku angielskim od sierpnia 2011 r. udostępniano w Internecie: http://bazy.wbp.krakow.pl/cgi-bin/makwww2//makwww.exe?BM=2&JE=A [29.08.2011].

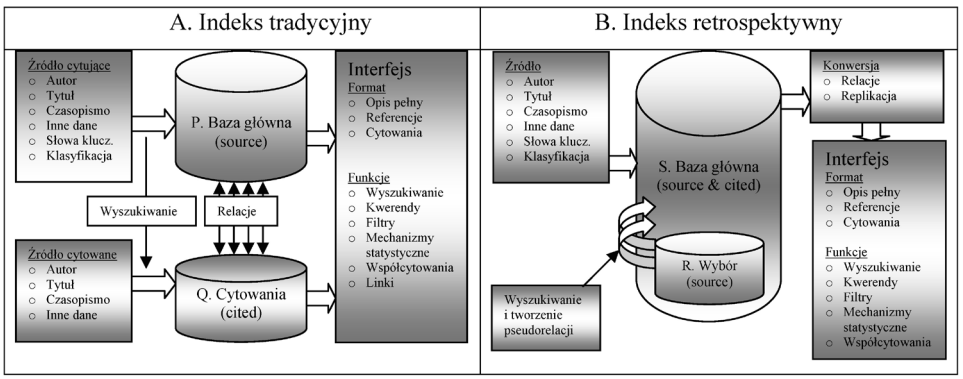
[5] Szczegółowe zasady doboru i selekcji opisane są w dalszej części niniejszego opracowania.

[6] ICHMP – liczby dot. tylko piśmiennictwa historycznego z lat 1945-2009; SCI 1981-1992 – źródło: (Kozłowski, 1994, s. 14-15); historia w A&HCI (Maršakova-Šajkevi, 2001, s. 157).

Głównym walorem tytułowego rozwiązania są: niskie koszty, duża skuteczność oraz ogromne możliwości obliczeniowe. Szczegóły metodologiczne, organizacyjne i techniczne zastosowane w ICHMP zostały opisane w kolejnych rozdziałach niniejszego artykułu.

MODEL I STRUKTURA INDEKSU CYTOWAŃ HISTORIOGRAFII MEDIÓW POLSKICH

Głównym problem stojącym przed projektantem indeksu cytowań jest ujęcie wszystkich procedur merytorycznych i organizacyjnych w ramach możliwości technicznych systemu informatycznego. Rozwiązania w tym zakresie mogą być rozmaite i zmieniać się wraz z rozwojem technologii. Przykładem jest SCI, który ewoluował od formy drukowanej (kartoteki), wspomaganej komputerowo tylko na etapie sortowania indeksów (lata 60.), poprzez uzupełnianą manualnie relacyjną bazę (od połowy lat 70.), do sytemu wspomaganego komputerowo (przełom XX/XXI w.). Jakkolwiek serwisy *Thomson ISI* i *Scopus* nie udostępniły publicznie szczegółów technicznych używanych rozwiązań, wiele danych można uzyskać z analizy danych, interfejsu lub literatury; najpełniejszy obraz tych procesów dał Eugene Garfield w szkicu *The Design and Production of a Citation Index*, zawartym w trzecim rozdziale jego monografii (Garfield, 1979). W świetle powyższych źródeł tradycyjny indeks cytowań można zobrazować, jako system składający się trzech elementów: wejścia (indeksowane źródła); relacyjnej bazy danych (tabele cytujące i cytowane) oraz interfejsu (rys. 1A).



Rys. 1. Schematy ideowe indeksów cytowań

Warto podkreślić, że w tak określonym schemacie występuje dwa rodzaje relacji: PQ (wszystkie pozycje cytowane [References]) oraz QP (te pozycje cytowane, które są jednocześnie cytującymi [Times cited]). Należy zwrócić uwagę, że jedynie te ostatnie są w pełni wykorzystane w rankingach bibliometrycznych. Według Janet Robertson, koordynatora projektów *Thomson ISI*, średni roczny przyrost cytowań PQ wynosi 25 mln, z czego jedynie 14 mln (56%) odnosi się do QP, czyli prac z *Master*

Journal List [MJL][7]. Oznacza to, że pozostałe rekordy (11 mln cytowań, czyli aż 44%) to pozycje nieistotne z punktu widzenia wskaźników bibliometrycznych. Powstaje pytanie, dlaczego ISI stosuje tak nieekonomiczną metodę? Procedura ta – jak przekonuje analiza – jest niezbędna, aby obsłużyć trzy typy zdarzeń: (1) właściwe przetwarzanie nowych rekordów Q (cited), które jeszcze nie zostały zarejestrowane w P (source) – powiązanie nastąpi w dopiero chwili, gdy zostanie on wprowadzony; (2) gromadzenie materiału o pozycjach często cytowanych, nienależących do P – jest to podstawą okresowych rewizji MJL (dodawanie nowych czasopism); (3) wspomaganie powiązań relacyjnych z użyciem słowników synonimów – dotyczy to powiązania błędnych rekordów ze zbioru Q (zwykle błędy pisowni, transliteracji, liczby i formy imion). Warto zwrócić uwagę, że wszystkie wskazane czynności odnoszą się wyłącznie do prac związanych z etapem bieżącego uzupełniania; nie mają natomiast wpływu na samo przetwarzanie już wprowadzonych danych.

Ostatni wniosek legł u podstaw domysłu, że do zbliżonych rezultatów może prowadzić baza pozbawiona tych mechanizmów. A zatem, jeśli zrezygnujemy z Q i jej miejsce zreplikujemy P można będzie tworzyć wyłącznie relacje P_pP_q i P_qP_p, czyli powiązania w obrębie jednej bazy. Schemat tego rozwiązania ilustruje rys. 1B, gdzie odpowiednikiem P_q+P_p jest zbiór S (*source* & *cited*). Warto dodać, że z merytorycznego punktu widzenia nie jest konieczne, aby relacje w obrębie zbioru głównego były wyczerpujące. W zupełności wystarczy wytypowanie i oznaczenie w ramach S podzbioru R, który będzie obejmował wybrane pozycje cytujące (*source*). A zatem do zrealizowania powyższej koncepcji można użyć dowolnej istniejącej bazy bibliograficznej. W praktyce możliwe są dwa rozwiązania: (1) zreplikowanie bazy i tworzenie relacji pomiędzy obiema replikacjami; (2) operowanie na jednej bazie poprzez tworzenie pseudorelacji, zaś replikację i generowanie relacji należy wykonać na końcu prac.

Drugie rozwiązanie, czyli tworzenie pseudorelacji jest łatwiejsze do wykonania, gdyż w praktyce sprowadza się do dwu czynności: (a) sygnowania rekordów cytujących unikalnym numerem oraz (b) dopisywanie powyższego numeru do każdego cytowanego przezeń rekordu w innym powtarzalnym podpolu (por. rys. 2). Ostatecznie cała procedura tworzenia retrospektywnego indeksu cytowań składa się z pięciu etapów:

1. Tworzenie bazy danych (S), które należy rozpocząć od doboru źródeł.
2. Typowanie publikacji cytujących (R).
3. Dopisywanie cytowań do bazy (czyli tworzenie pseudorelacji).
4. Konwersja danych, w wyniku której z pseudorealcji generowany jest tzw. klucz obcy oraz zostaje zreplikowana baza główna.
5. Eksploatacja – uzyskane w punkcie 4 komponenty można już przetwarzać w dowolnym relacyjnym systemie obsługi baz danych (np. MySQL, Access itp.).

TWORZENIE BAZY. ZASIĘG. ZAKRES. OPIS. JEDNOSTKI

Tworzenie bazy przeznaczonej do obsługi cytowań nie różni się zasadniczo od tworzenia bazy dziedzinowej, stąd i metoda jest bardzo zbliżona.

[7] *Cited Title Unification* – http://thomsonreuters.com/products_services/science/free/essays/cited_title_unification/ [2011.06.11].

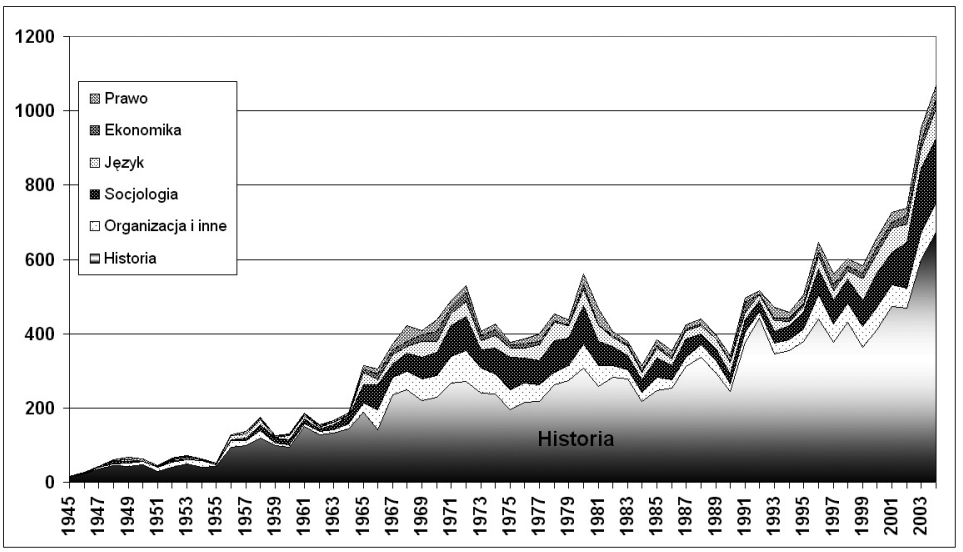
Wynikają z tego dwie konsekwencje: możliwe jest wykorzystanie do tego celu dowolnej istniejącej bazy bibliograficznej oraz znajdą tu zastosowanie zasady tradycyjnej metodyki bibliograficznej (zasięg, zakres, opis, jednostki). Istnieją wprawdzie pewne różnice między obiema formami (o których niżej), ale zasadniczo sprowadzają się one do nieznacznych korekt na istniejących rekordach.

W humanistyce okres starzenia się publikacji jest dość długi, stąd najkorzystniej jest nie stosować w bazie cytowań żadnych ograniczeń chronologicznych, gdyż prowadzi to do deformacji obrazu dyscypliny. Przykładowo w literaturoznawstwie aż 46,7% cytowań wskazuje na pozycje liczące więcej niż 16 lat, w tym 17,0% – na starsze niż 50 lat (Konieczna, 2002); w historii zaś odpowiednio: 56,61% i 7%; podczas gdy w naukach przyrodniczych dzieje się zgoła odwrotnie – np. cytaty młodsze niż 10 lat stanowią aż 88,2% odwołań w fizyce i 71,2% w chemii (Stern, 1983). Z drugiej strony zarejestrowanie kompletu piśmiennictwa z najmniejszej nawet dziedziny wydaje się niemożliwe do wykonania. Wyjściem z sytuacji jest zastosowanie dwu zabiegów: (a) selekcji – ograniczającej zbiór do pozycji ściśle naukowych; (b) ewentualne ograniczenie zasięgu chronologicznego do daty instytucjonalizacji badanej nauki (zwykle jest to data rozpoczęcia publikacji dziedzinowego czasopisma naukowego); w przypadku zastosowania innej cezury starsze publikacje będą miały zaniżone wskaźniki. Nie jest natomiast wskazane stosowanie ograniczeń w zakresie form wydawniczych, gdyż w humanistyce wszystkie są często używane. Warto jednak zaznaczyć, że szczególnie pieczołowicie należy traktować wydawnictwa zwarte, gdyż jakkolwiek stanowią one kilkanaście procent ogólnej liczby jednostek, to są jednocześnie grupą, która generuje i gromadzi najwięcej cytowań, np. w historii książki stanowią aż 92% prac mających co najmniej po 50 cytowań oraz 53% – co najmniej po 10.

Nie wydaje się także celowe ograniczanie zasięgu terytorialnego lub językowego. Wprawdzie badania dowodzą, że większość dyscyplin humanistycznych jest lokalna językowo i nie należy się spodziewać, by pozycje o doniosłej wartości ukazywały się w językach kongresowych (jeśli powstają, to głównie w celach popularyzatorskich); istnieją jednak wyjątki, które należy tłumaczyć specyfiką niektórych dyscyplin (np. archeologia, filologia klasyczna, które w dużej mierze funkcjonują w obiegu międzynarodowym) lub w przypadku innych dyscyplin – czynnikami historycznymi. Przykładami takich wyjątków w obszarze objętym problematyką bazy ICHMP (czyli w historii) są prace dotyczące życia i kultury emigracji, badania żydowskiej diaspory lub pogranicza polsko-niemieckiego. Tym samym powodem uzasadniany jest też szeroki zasięg terytorialny omawianej bazy. W tym przypadku chodzi o silne środowiska emigracyjne (Paryż, Londyn), gdzie wydano szereg cenionych i cytowanych prac (np. w kręgu paryskiej „Kultury"). Inne przypadki wychodzące poza ten krąg stanowią margines. Także badania polskiej humanistyki (np. historii i filologii) podejmowane przez uczonych zagranicznych są rzadkie i najczęściej dotyczą komparatystyki międzykulturowej.

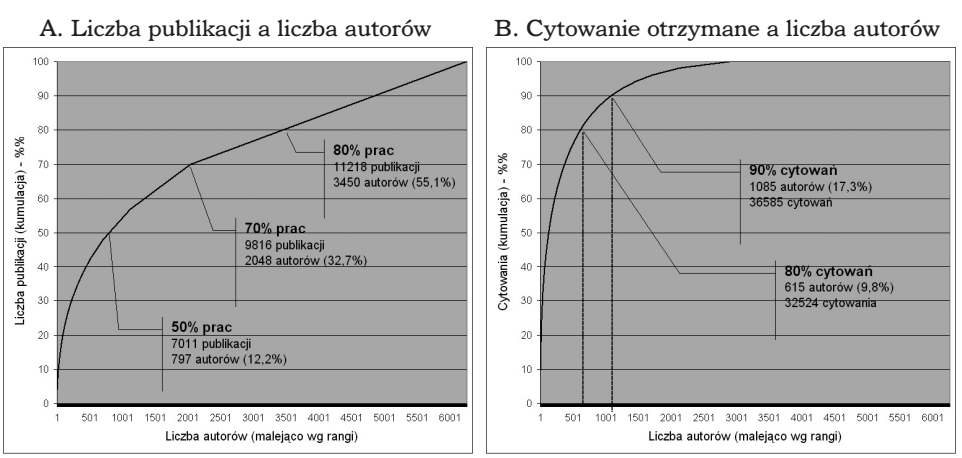
Więcej problemów sprawia właściwy dobór zakresu tematycznego. Na potrzeby indeksu cytowań precyzyjne zdefiniowanie i ścisłe ograniczenie się do wyznaczonego pola nie wystarcza, gdyż cytowania ujawniają powiązania międzydyscyplinarne. Słowem, najkorzystniej jest w bazie uwzględnić pewien materiał kontekstowy (z dyscyplin pogranicznych i pokrewnych).

Podkreślmy jednak, że jeśli zabraknie takiego otoczenia nie odbije się to szczególnie niekorzystnie na wynikach, gdyż w humanistyce najważniejsze prace powstają na macierzystych polach badawczych, czyli w ramach dyskursu właściwego dla danej dyscypliny. Poświadczają to badania z tego zakresu, np. w literaturoznawstwie większość cytowań (64,1%) odwoływało się do macierzystego korpusu, zaś pozostałe 35,8% na dyscypliny stanowiące kontekst (sztuka, historia, kultura i in.) (Konieczna, 2002). W ICHMP uzyskano silniej spolaryzowane dane, gdyż na publikacje z zakresu historii mediów powoływało się aż 87,7% prac z własnego korpusu, pozostałe 12,3% rozkładało się na: socjologię mediów – 5,2%; język mediów – 1,7%; prawo mediów – 1,2%, ekonomikę mediów – 0,5% oraz inne – 3,5%. Można więc zaryzykować tezę, że dla humanistyki bliższy prawdy jest ostatni wskaźnik (87,7%), gdyż w ICHMP materiał był dobrany bardzo starannie i reprezentatywnie (wykr. 1).



Wykres 1. Zawartość Indeksu Cytowań Historiografii Mediów Polskich [liczba publikacji]

Istotną rolę w budowie indeksu odgrywają procesy uzupełniania i selekcji, należy je jednak właściwie interpretować i odpowiednio rozłożyć akcenty. Kluczową kwestią jest włączenie do bazy kompletu prac istotnych dla badanej dyscypliny. Jeśli zauważymy, że z naukoznawczego punktu widzenia istotne są jedynie publikacje o charakterze ściśle naukowym, szczególnie te, które zawierają oryginalne wyniki badań, łatwo skonstatować, że ważna jest rejestracja dorobku naukowców z tzw. grupy podstawowej. Badania przekonują, że w każdej dyscyplinie istnieje taka grupa, która jest zarówno płodna, jak i wysoko cytowana. O konsekwencjach i sile tego związku przekonują poniższe wykresy, z których wynika, że w historii mediów 12,2% najpłodniejszych autorów dostarcza 50% prac (wykr. 2A), zaś 9,8% najczęściej cytowanych uczonych skupia aż 80% cytowań (wykr. 2B); warto dodać, że istnieją przekonujące dowody, że obie grupy pokrywają się w znacznym stopniu (korelacja R=0,78) (Kolasa, 2011b).



Wykres 2. Ranga autora a produktywność i cytowania w historii mediów polskich [%%] (Kolasa, 2011a)

W ślad za tym najważniejsza jest identyfikacja i dokładna rejestracja prac autorów ściśle związanych z badaną dyscypliną. A zatem pierwszym, najważniejszym krokiem przy tworzeniu indeksu jest zebranie względnie kompletnego dorobku tej grupy. W przypadku pozostałych prac warto zastosować selekcję, aby wytypować tylko publikacje o charakterze ściśle naukowym i prace dokumentacyjne o trwałej wartości. Jakkolwiek w indeksie cytowań można zrezygnować z tego kroku, gdyż publikacje mało znaczące – po prostu nie będą cytowane, brak selekcji będzie jednak skutkować koniecznością wprowadzenia ogromnej liczby peryferyjnych rekordów, co pociągnie za sobą wiele niepotrzebnej pracy. Selekcję powinna wykonać osoba znająca badaną dyscyplinę i jej środowisko naukowe, najlepiej czynny uczony. Ogromną pomocą w tej pracy mogą być przeglądy dokumentacyjne lub bibliografie abstraktowe. W ICHMP rolę taką spełniły *Polska bibliografia adnotowana wiedzy o środkach masowego komunikowania* (1965-1987) oraz *Polska bibliografia prasoznawcza 1996-2001* (Kolasa & Jarowiecki, 2005), które uzupełniono informacjami z 34 innych źródeł (bibliografii historycznych, bibliologicznych, prasoznawczych) zarówno polskich, jak i zagranicznych (np. *International Bibliography of Historical Sciences* 1926-2003).

Opis bibliograficzny stosowany w bazach cytowań zwykle różni się poziomem precyzji i doborem elementów od opisu używanego w bazach bibliograficznych. Różnice te wynikają jednak przede wszystkim z uwarunkowań historycznych. Nie istnieją żadne przeszkody, aby do tego celu użyć standardu MARC, uzupełnionego o elementy istotne w przetwarzaniu. Zastosowanie MARC przynosi natomiast wiele korzyści, w szczególności zwiększa poziom precyzji danych i stwarza możliwość importu rekordów z innych źródeł. Niedoskonałości MARC ważne z punktu widzenia przetwarzania i statystyk (szczególnie znaki umowne) można bezstratnie wyeliminować na etapie konwersji. Z drugiej strony do kontroli tych elementów, które mają kluczowe znaczenie w obliczeniach należy bezwzględnie zastosować kartoteki haseł wzorcowych. Dotyczy to szczególnie pól przeznaczonych na nazwy osobowe i skróty tytułów czasopism; podobnie należy kontrolować dane z pól danych kodowanych MARC oraz pól do-

datkowych (np. typ publikacji, zasięg, klasyfikacja). Zlekceważenie tego kroku sprawi, że podczas eksploatacji wyniki mogą ulec rozproszeniu.

Ważną kwestią przy budowie indeksu cytowań jest poprawne zdefiniowanie jednostek bibliograficznych. Przedmiotem cytowań powinny być całości samoistne piśmienniczo (artykuły, rozprawy z książek, broszury, książki), można też wprowadzać fragmenty książek. Nie należy natomiast odrębnie rejestrować recenzji i polemik, lecz dopisywać je do pozycji, do których się odnoszą. Recenzje są bowiem publikacjami ważnymi jedynie w aspekcie krytyki naukowej, ale bezużyteczne z informacyjnego punktu widzenia. Dodajmy, że badania na bazie ICHMP (z użyciem 3054 recenzji) wykazały, że nie istnieje żaden racjonalny związek pomiędzy częstością recenzowania a cytowaniem. Druga uwaga odnosi się do konieczności dublowania niektórych opisów. Obserwacja praktyki wskazuje, że w odniesieniu do wydawnictw zbiorowych istnieje duża dowolność w obyczajach cytowania (część wskazuje na całość, inni na zawarte tam prace). Aby wyeliminować tę wieloznaczność wydawnictwa tego rodzaju należy opisać podwójnie: zarówno każdą zawartą pracę, jak i całość wydawnictwa (czyli na dwóch poziomach z zawartością w adnotacji). Uwaga ta dotyczy jedynie prac zbiorowych ściśle związanych z badaną dyscypliną.

Zarówno proces tworzenia bazy, jak i procedura dopisywania cytowań wymaga wykonania ogromnej liczby powtarzalnych czynności, które z racjonalnych względów warto rozłożyć na wiele osób. Aby dane były integralne i nie dochodziło do kolizji cały proces, należy przeprowadzać na centralnej bazie z obsługą kartotek wzorcowych w systemie wielodostępnym o definiowanych uprawnieniach. Innymi słowy, system taki powinien być zorganizowany w sposób zbliżony do NUKat. W ICHMP wykorzystano do tego celu edukacyjną instalację systemu SOWA2/SQL (PostgreSQL) udostępnianą z serwera producenta. Skalę obciążenia ilustrują liczby: w rozmaitych okresach w latach 2005-2009 na bazie pracowało ok. 500 osób, które razem wykonały ok. 170 tys. modyfikacji zbioru liczącego ok. 24 tys. rekordów, zatem każdy rekord był modyfikowany średnio 7 razy, z czego 4 razy na etapie tworzenia i korekty i 3 razy w fazie dodawania cytowań.

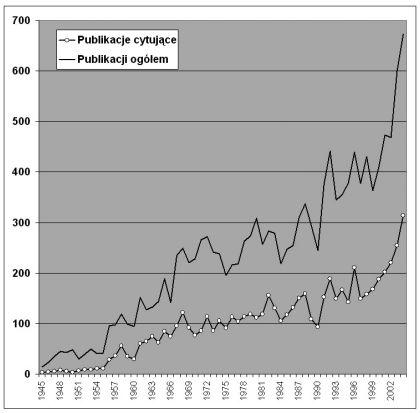
TYPOWANIE POZYCJI CYTUJĄCYCH. TWORZENIE CYTOWAŃ

Procedurę tworzenia cytowań w indeksie retrospektywnym należy poprzedzić wytypowaniem pozycji cytujących, czyli zdefiniować zbiór R (por. rys. 1B). Pomocne w tym zakresie winno okazać się prawo S. C. Bradforda (Nowak, 2008, s. 69-74). Aby zmaksymalizować efekty, należy zatem najpierw wprowadzić artykuły zawarte w czasopismach z rdzenia dyscypliny, a następnie stopniowo eksploatować tytuły z dalszych grup. W ICHMP proces ten przebiegał modelowo, gdyż w całym badanym okresie ukazywało się jedno pismo poświęcone ściśle badanej tematyce[8]. Ostatecznie grupa podstawowa dostarczyła 14,3% pozycji cytujących, zaś kolejne grupy (zawierające podobną liczbę artykułów) liczyły: 17, 100 i 525 tytułów czasopism. Koncentracja jest łatwo uchwytna, lecz trudno w wyliczeniu dopatrzeć się klasycznej formuły Bradforda ($n^0 : n^1 : n^2 : \ldots$

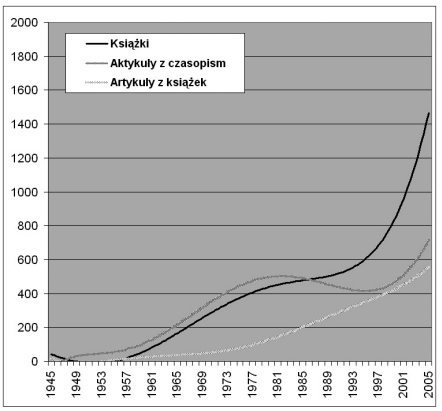
[8] „Rocznik Historii Czasopiśmiennictwa Polskiego" (1962-1976); nast. „Kwartalnik Historii Prasy Polskiej" (1977-1993); nieformalnie kontynuowany przez „Rocznik Historii Prasy Polskiej" (od 1998).

: n^n). Wynika to po pierwsze z faktu, że artykuły z czasopism stanowiły jedynie 55,4% pozycji cytujących; po wtóre – grupa trzecia i dalsze były poddane selekcji, po trzecie – nie analizowano korpusu czasopism ukazujących się równolegle. Ważniejsza okazuje się jednak konstatacja, że czasopisma wyprodukowały jedynie 32,3% cytowań. Teza ta ujawnia ważną cechę humanistyki: jakkolwiek wydawnictwa zwarte stanowiły jedynie 12,7% prac cytujących, lecz pochodziło z nich aż 47,2% cytowań; zaś artykuły zawarte książkach (pracach zbiorowych) – odpowiednio 29,4 i 17,5 (wykr. 3B); prawie nieistotne okazały się natomiast inne formy (np. broszury, fragmenty) – 2,4 i 2,8. Wykaz udowadnia kontrowersyjną tezę, że wśród priorytetów w procesie typowania należy w pierwszej kolejności uwzględnić wszystkie wydawnictwa zwarte, a dopiero potem stosować prawo Bradforda.

A. Prace cytujące na tle ogółu publikacji



B. Produkcja cytowań wg norm wydawniczych



Wykres 3. Wybrane parametry ICHMP

Problem dalszej strategii typowania pojawi się nieuchronnie po wyczerpaniu się listy książek i głównych tytułów czasopism, w szczególności przy doborze artykułów z książek. W ICHMP wykorzystano do tego celu same cytowania. Procedura była nieskomplikowana: po wprowadzeniu ok. 75% cytowań wygenerowano listę pozycji najwyżej cytowanych, które do tej pory nie były wytypowane; następnie dodano je do listy pozycji cytujących i wyszukiwano cytowania. Procedurę tę powtarzano następnie kilkakrotnie aż osiągnięto satysfakcjonujące współczynniki. W ten sposób indeks stał się systemem samoorganizującym i zaczął pełnić funkcje semantyczne w procedurze typowania. Ostatecznie więc każda pozycja wysoko cytowana była jednocześnie jednostką cytującą, zaś w grupie prac rzadziej cytowanych – tylko te były jednostkami cytującymi, które dobrano w pierwszym kroku (książki, artykuły z głównych czasopism i prac zbiorowych). Taka taktyka – warto zauważyć – była wyjściem naprzeciw tezie, że cytowania wyprodukowane (użyte) przez autorów uznanych (wysoko cytowanych) mają większą wartość niż cytowania pozyskane z publikacji okazjonalnych. Teza ta nie wymaga dodatkowego uzasadnienia, gdyż leży niejako u podłoża samej idei cytowania (Garfield, 1955). Dodajmy, że podczas typowania wraz z kolejnym poszerzaniem listy efekt przyrostu cytowań stawał się coraz mniejszy i ostatecznie zaprzestano typowania

po osiągnięciu 6880 prac cytujących; stanowiło to 43% zawartości bazy, które rozłożyły się proporcjonalnie do ogólnej liczby prac na osi czasu (wykr. 3A). Z drugiej strony warto zauważyć, że zależność pomiędzy liczbą pozycji cytujących uszeregowanych malejąco wg rangi autora jest krzywą logistyczną o charakterystyce podobnej do krzywej z wykresu 2A.

Techniczna realizacja procedury cytowania jest prosta i sprowadza się do dwóch kroków: (1) sygnowania rekordów cytujących unikalnym numerem w odpowiednim niepowtarzalnym podpolu; (2) dopisywania powyższego numeru do każdego cytowanego przezeń rekordu w innym, powtarzalnym podpolu. Czynność tę ilustruje poniższy przykład (rys. 2), gdzie numer cytujący został wpisany do pola 024%a, zaś cytowania do powtarzalnego pola 033%a (Kolasa, 2009).

```
LDR   %b n %c a %d s %e + %f + %g + %h +450+
001   %a 0000654332
005   %a 20100618013633.0
008   %a 080830 %l s %m 2004 %n ++++ %b ### %h #### %c # %d r %p #### %r # %s 0
      %f 0 %t # %u + %g # %w # %j pol %i # %e #
024   %a 6329
033   %a 2906 %a 2910 %a 3453 %a 3592 %a 3409 %a 5761 %a 3530 %a 7824 %a
          7902 %a 8510 %a 8526
      %a 8535 %a 8641 %a 6315 %a 6470 %a 6469 %a 6561 %a 6594 %a 6619 %a
          6695 %a 6759
040   %a wmk %c wmk %d wmk
0410  %a pol
1001  %a Kolasa, Władysław Marek
24510 %a Prasa Krakowa w dekadzie przemian 1989-1998 : %b rynek, polityka, kultura /
      % c Władysław Marek Kolasa.
260   %a Kraków : %b Wydaw. Naukowe AP, %c 2004.
....
```

Rys. 2. Rekord cytujący i cytowany w ICHM

Bez wątpienia najbardziej pracochłonnym etapem pracy nad bazą jest proces dopisywania cytowań. Skala problemu ujawnia się już na poziomie szacunków: przy założeniu, że przeciętny artykuł w humanistyce ma ok. 19 odwołań w przypisach lub bibliografii (Konieczna, 2002), zaś książka 5-6 razy więcej (114) należy w zbiorze 6880 pozycji (877 książek i 6003 artykułów) spodziewać się 214 035 cytowań. Oznacza to konieczność bezbłędnego wpisania do interfejsu wyszukiwawczego prawie ćwierć miliona poszukiwanych fraz. Obserwacja ta skłoniła do szukania ułatwień i częściowej automatyzacji tego procesu. W ICHMP każdy artykuł wytypowany do cytowania był najpierw zamawiany w bibliotece, następnie wykonywano dobrej jakości kserokopię; ta zaś była znakowana numerem cytującym (024%a) i poddana digitalizacji. Łącznie zdigitalizowano 6003 artykułów (razem 101 721 stron), a sam proces skanowania (średnio 20 sekund na stronę przy 600 dpi) trwał 561 godzin. Skany w dużych partiach były okresowo poddawane procesowi OCR (z użyciem Abbyy FineReader), następnie zapisywane jako hybrydowy PDF. Dodatkowym efektem procesu stała się pełnotekstowa baza wszystkich artykułów. Nieco inaczej postępowano z książkami (razem 249 201 stron) – skanowano jedynie fragmenty zawierające bibliografię załącznikową. Testy udowod-

niły bardzo wysoką efektywność tej techniki, gdyż skuteczność OCR dla większości druków przekraczała ok. 95%, co pozwoliło zrezygnować z ręcznego wpisywania referencji podczas wyszukiwania. W efekcie proces dopisywania cytowań został zredukowany do sekwencji „kopiuj – wklej", zaś pojedyncza operacja trwała ok. 3-7 sekund (gdy brak pozycji) lub 10-30 sekund (gdy pozycję znaleziono i dopisano numer).

Warto zauważyć, że w indeksie retrospektywnym dopisywane są jedynie cytowania odwołujące się do pozycji zarejestrowanych, stąd ich liczba jest mniejsza niż rzeczywista liczba odwołań. W ICHMP uzyskano 63 811 cytowań, z czego najwięcej pochodziło z książek – średnio po 32,2 referencji na jednostkę, podczas gdy artykuły generowały średnio po 5,1. Uzyskana liczba (63811) stanowi ok. 25% cytowań potencjalnych, co oznacza, że osiągnięto o połowę niższą skuteczność od baz ISI (56%); nie wynika to jednak z błędów metodologicznych, lecz specyfiki pisarstwa humanistycznego (historycznego w szczególności). W dziedzinach tych znacząca liczba referencji nie odwołuje się do literatury przedmiotu, lecz rozproszonych materiałów źródłowych (np. doniesień prasowych). Warto dodać, że w trakcie prac nad ICHMP stwierdzono, że średnio co szósty przypis z cytowanych prac zawierał błędy lub pomyłki (16%), które jednak przy sprawnie działających mechanizmach wyszukiwawczych można było zidentyfikować i stworzyć poprawne cytowanie.

W teorii można rozważać dalej idące uproszczenia w opisanej technice. W szczególności rozważyć zastosowanie automatycznego wyszukiwania przypisów (referencji) wspomaganych bazą danych. Prace koncepcyjne w tym zakresie osiągnęły już wysoki poziom zaawansowania i znalazły zastosowania w praktyce; np. metoda *autonomous citation matching* stosowana w serwisie *CiteSeer*, którego skuteczność ocenia się na 80-90% (Giles et al., 1999; Lawrence et al., 1999), czy rozwiązania platformy *CitHit* doskonalone na potrzeby *Medline*[9], a w szczególności projekt *GROBID* (Lopez, 2009). Bliższe testy udowodniły jednak, że rozwiązania te sprawdzają się jedynie w publikacjach stosujących ściśle reżimy cytowania (zwykle formy dedykowane, np. przyjęte w *Medline*); w efekcie dotychczas nie zastosowano ich nawet u głównego dostawcy cytowań *Thomson ISI*. Tym bardziej trudno rozważać ich wykorzystanie w odniesieniu do humanistyki, gdzie sposób prezentacji cytowanej literatury jest bardzo zindywidualizowany. Wciąż formą dominującą są tu obszerne przypisy umieszczone pod tekstem strony, nierzadko złączone składniowo z komentarzem i autorskim skrótami przeplecione znaczącą liczbą powtórzeń (*op. cit.* itp.). W porównaniu ze standardami stosowanymi w naukach ścisłych (style: Harvard, APA, MLA, Turbian) technika taka jest bardzo niewygodna na etapie indeksacji i utrudnia automatyzację.

KONWERSJA

Ostatnim etapem budowy indeksu jest konwersja, dzięki której dotychczasowa baza z pseudorelacjami zostanie przekształcona do postaci nadającej się do eksploatacji w dowolnym systemie obsługi relacyjnej bazy danych, np. MySQL, Access itp. Procedura składa się z czterech

[9] *CitHit Citation Recognition System*: http://www.cs.uwm.edu/~qing/projects/cithit/index.html [2011.05.21].

etapów: (1) eksport danych do postaci wymiennej (najlepiej ISO2709); (2) import do systemu przejściowego, wygenerowanie klucza i wykonanie ewentualnych korekt grupowych; (3) eksport odpowiednich tabel do postaci wymiennej (najlepiej tekst rozdzielany); (4) import do systemu obsługi bazy relacyjnej (DBMS), replikacja tabel i eksploatacja.

Pierwszym krokiem jest eksport danych MARC do postaci ISO2709 i import do systemu przejściowego (np. MAK), aby przygotować dane do utworzenia odpowiednich tabel. Należy w szczególności przygotować dane do tabeli głównej (tbMaster), która może zawierać wszystkie pola lub tylko te, na których będą wykonywane obliczenia. Na tym etapie należy nadać rekordom unikalne numery; sprawdzić poprawność danych oraz wykonać ewentualne modyfikacje grupowe (np. redukcję znaków umownych MARC). Gdy dane są gotowe należy wybrać w masce odpowiednie pola i wyeksportować w wymiennym formacie tekstowym (w MAK-u opcja: IMPEK/relacyjne). Ważną czynnością jest przygotowanie klucza, który uzyskujemy z przekształcenia pseudorelacji. W tym celu należy wykonać trzy kroki: (1) zmienić wielokrotne wystąpienia podpola 033%a na wielokrotne wystąpienie pola 033; (2) dopisać do każdego pola 033 podpole z numerem sekwencyjnym; (3) wyeksportować wszystkie pola 033 jako tekst – efekt ilustruje rysunek 3.

Tabela 3

Przekształcenie pseudorelacji w klucz obcy

Etap 1	Etap 2	Etap 3
001 %a 0000654332 024 %a 6329 033 %a 2906 033 %a 2910 … 001 %a 0000784336 033 %a 6470	001 %a 0000654332 024 %a 6329 033 %n 0000654332 %a 2906 033 %n 0000654332 %a 2910 … 001 %a 0000784336 033 %n 0000784336 %a 6470	0000654332; 2906 0000654332; 2910 … 0000784336; 6470 …

Po przygotowaniu danych do wszystkich tabel można je zaimportować do dowolnego systemu obsługi bazy danych i przetwarzać z użyciem dostępnych tam narządzi. Duże znaczenie dla większości kwerend ma jednak właściwe połączenie komponentów. W najprostszej postaci potrzebne są trzy tabele: tbMaster (tabela gówna), tbKey (klucz obcy) i tbCiting (tabela cytująca, czyli kopia tbMaster), które należy połączyć relacjami zgodnie z rys. 4[10].

[10] Warto zauważyć, że chociaż klucz obcy z tabelami jest sprzężony relacją *jeden do jednego* w istocie odwzorowuje relację *wiele do wielu*. W przypadku stwierdzenia braku jakiś danych w tabeli można je stosunkowo łatwo dodać; wystarczy przygotować odpowiednie dane jako oddzielna tabela i przyłączyć do w relacji 1:1 za pomocą pola 001 do dowolnej z tabel; należy tak postąpić w szczególności z polami powtarzalnymi (np. 7xx, 6xx itd.).

Tabela 4

Podstawowe połączenia relacyjne dla kwerendy przetwarzającej cytowania

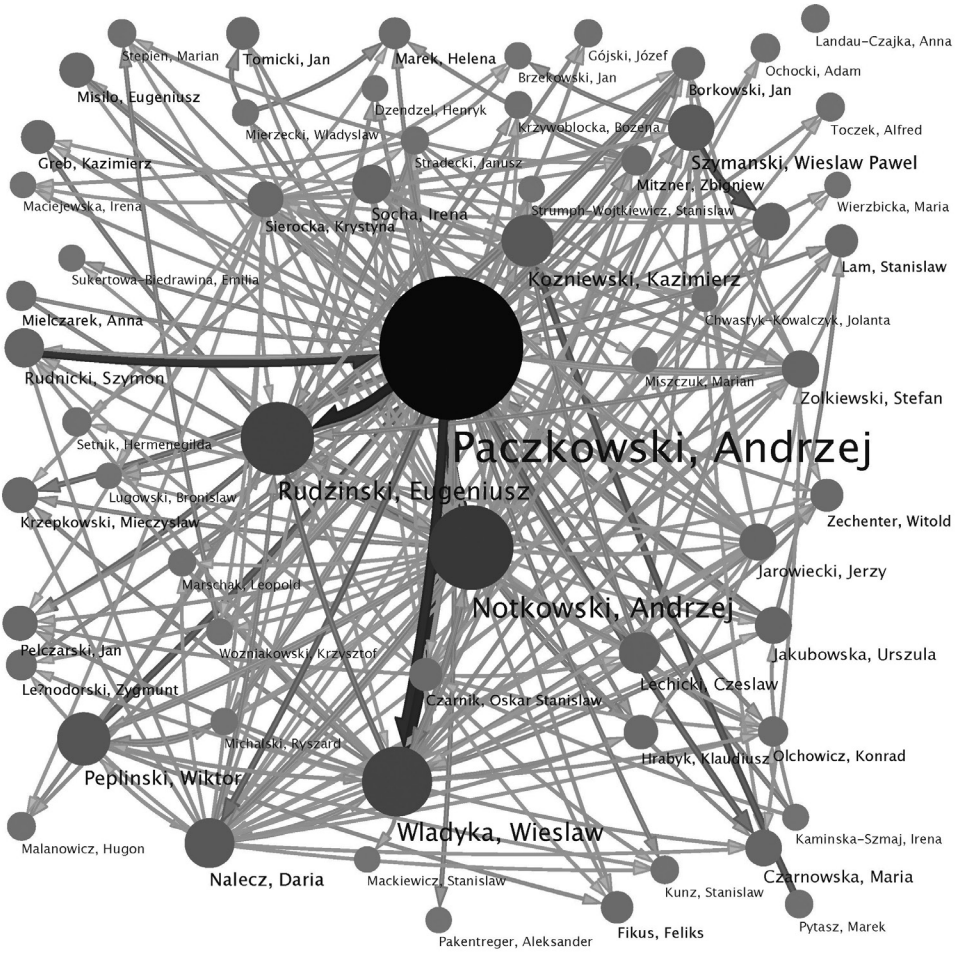
tbCiting (cytująca)					Relacja	**tbKey** (klucz)		Relacja	**tbMaster** (cytowana)				
…	Rok	Tytuł	Autor_1	024	1:1	033	001	1:1	001	Autor_1	Tytuł	Rok	…
…	2005	Prasa…	Wiatr, Emil	2960		2960	565665		565665	Rusek, Filip	Prasa…	2004	…
…	2005	Prasa…	Wiatr, Emil	2960		2960	785663		785663	Maj, Józef	Media…	2001	…

Na postawie istniejących komponentów można, zależnie od potrzeb, definiować odpowiednie kwerendy (wybierające, krzyżowe, parametryczne itd.) i przeprowadzać odpowiednie obliczenia. Dane w tej postaci pozwalają na obliczanie dowolnych zależności bibliometrycznych, np. cytowań czy rozkładów dowolnej cechy (zwykłe, krzyżowe i warunkowe), a więc umożliwiają pełny dostęp do zależności pomiędzy zbiorem cytującym a cytowanym. Zaletą zaprezentowanego rozwiązania jest też możliwość definiowania rozmaitych warunków umożliwiających korekty wskaźników, np. eliminację autocytowań lub wybór cytowań o określonej aktualności. W większości sytuacji pomocny będzie wbudowany do bazy język skryptowy lub SQL oraz znajomość dostępnych tam funkcji, np. w pierwszym przypadku sprawdzi się funkcja *InStr*: ((InStr([tbCiting]![Autor_1],[tbMaster]![Autor_1]))=0 Or (InStr([tbCiting]![autor_1],[tbMaster]![autor_1])) Is Null)); z kolei dla szukania cytowań nie starszych niż np. 5 lat zwykła zależność arytmetyczna: (([tbCiting.Rok]-[tbMaster.Rok]))<=5)). Nie wszystkie istotne zależności można wykonać w ramach możliwości systemu obsługi bazy danych i języka SQL (np. liczenie wariancji, korelacji itp.). W takim przypadku wygodniej jest uzyskane tabele zaimportować do arkusza kalkulacyjnego. Także część złożonych funkcjonalności, jak liczenie wartości H-index, czy tworzenie kluczy dla współcytowań należy użyć zewnętrznych programów lub skryptów.

WYBRANE ZASTOSOWANIA

Dane uzyskane z indeksu cytowań – jak wspomniano na wstępie – pozwalają na rozmaite ich wykorzystanie. Warto jednak zwrócić uwagę na kilka niekonwencjonalnych obszarów, w szczególności wykorzystania cytowań do obrazowania trendów rozwojowych nauki. W dotychczasowej praktyce metody takie stosowano w bardzo wąskim zakresie, w zasadzie wyłącznie na użytek map i atlasów nauki ISI (Marshakova-Shaikevich, 1996, s 115-127). W ostatnich latach za sprawą nowych narzędzi zapożyczonych z teorii sieci społecznych obszar ich stosowania znacznie się poszerzył; przykładem są programy dedykowane, np. *HistCite* przeznaczony do operowania na danych ISI (Garfield, 2004) czy *Map Generator* z serwisu *Scimago*[11] działający w oparciu o dane *Scopusa*. Powstało też wiele programów umożliwiających wizualizację cytowań na dowolnych danych, m.in.: *Pajek*[12], *Visione*[13], czy *Map Equation*[14]. Ilustracją ich zastosowania jest wygenerowany na bazie danych ICHMP histogram (rys. 4), który przedstawia głównych badaczy prasy polskiej okresu II RP wraz z siecią łączących ich zależności.

[11] *Scimago/Map Generator:* http://www.scimagojr.com/mapgen.php [2011.06.20].
[12] *Pajek Wiki:* http://pajek.imfm.si/doku.php [2011.06.20].
[13] *Visione:* http://visone.info/doku.php [2011.06.20].
[14] *Map Equation:* http://www.mapequation.org [2011.06.20].

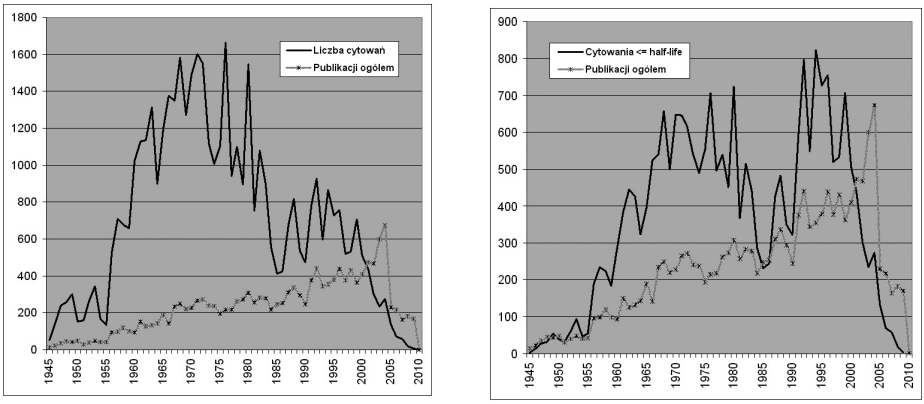


Rys. 4. Główni historycy prasy polskiej okresu 1918-1939 (współcytowania)

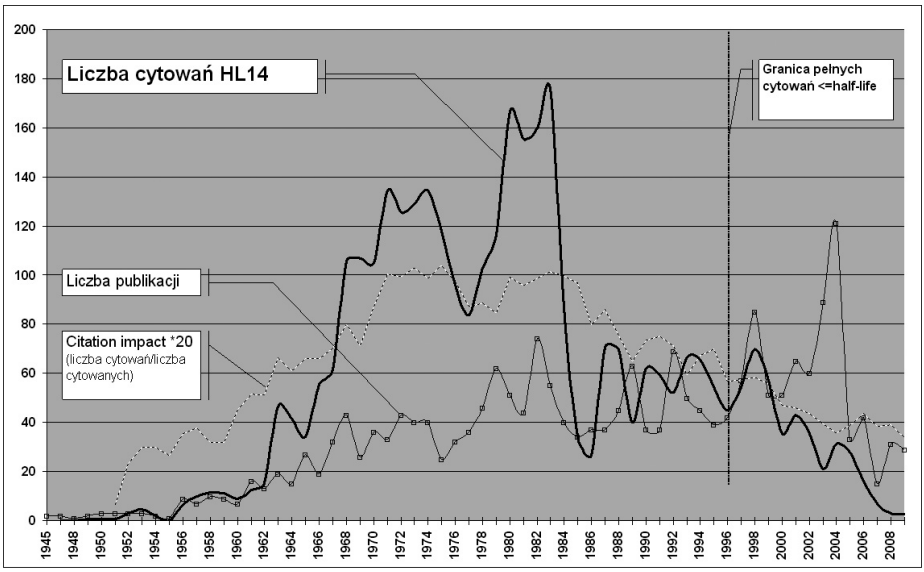
Bardziej użyteczną formą prezentacji rozwoju dyscypliny są wykresy obrazujące dynamikę rozwoju poszczególnych pól badawczych na osi czasu (wykr. 5). Warto jednak zaznaczyć, że proste zliczanie pełnej liczby cytowań deformuje obraz dyscypliny, gdyż faworyzuje publikacje starsze (wykr. 4A). Aby wyeliminować ten niekorzystny efekt, należy przy tego typu analizach operować skorygowaną liczbą cytowań. Badania dowiodły, że można do tego celu wykorzystać cytowania młodsze od *half-life* [HL14] (wykr. 4B).

W tak określonym modelu teoretycznym poprawna wizualizacja rozwoju dowolnego pola badawczego jest nieskomplikowana; obrazuje ją kolejny wykres, w którym zestawiono główne parametry analizowanego obszaru na osi czasu (wykr. 5).

A. Liczba publikacji i cytowania [100%] B. Liczba publikacji i cytowania skorygowane [HL14]



Wykres 4. Wybrane zależności dynamiczne w historiografii prasy (na bazie ICHMP)

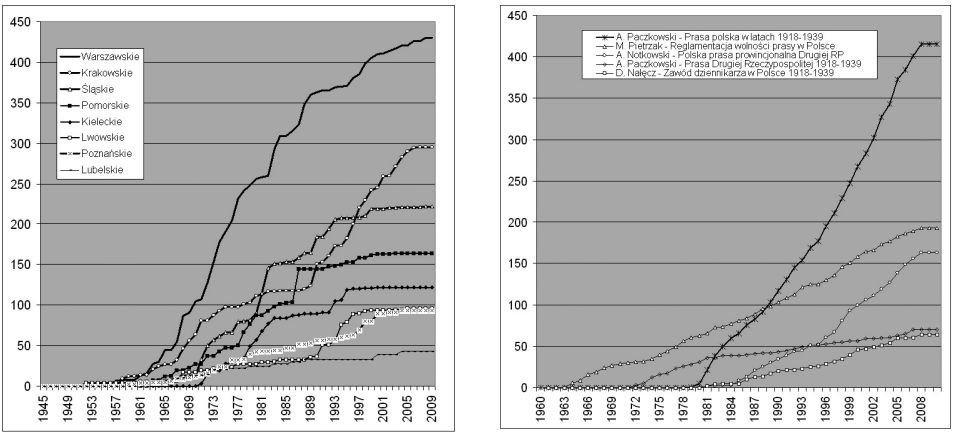


Wykres 5. Dynamika badań nad prasą polską okresu 1918-1939 w latach 1945-2009

Przedstawiony na wykresie 5 obraz rozwoju pola badawczego (dynamikę zmian reprezentuje krzywa cytowań HL14) precyzyjnie obrazuje okresy wzrostu, regresu oraz stagnacji i może z powodzeniem służyć jako podstawa do budowania narracji historiograficznej. Jedyną wadą tej metody wizualizacji jest granica pełnych cytowań, czyli miejsce, powyżej którego cytowania są niepełne, gdyż nie minął jeszcze pełny czas *half-life* (w przykładzie 14 lat).

W analogiczny sposób można też obrazować inne zmienne badanego korpusu nauk na osi czasu, np. cytowania pojedynczego uczonego, zestawiać dynamikę kilku pól badawczych (wykr. 6A) lub badać recepcję poszczególnych dzieł (wykr. 6B).

A. Geografia prasy polskiej [cytowania HL14] B. Recepcja głównych syntez [cytowania całkowite]



Wykres 6. Wybrane trendy rozwojowe w badaniach nad prasa polską okresu 1918-1939

LITERATURA

Anzenbacher, Arno (1992). *Wprowadzenie do filozofii*. Wyd. 2. Kraków: UNUM.

Archambault, Éric; Gagné, Étienne Vignola (2004). *The Use of Bibliometrics in the Social Sciences and Humanities*. Montreal: Science-Metrix.

Bakkalbasi, Nisa; Bauer, Kathleen; Glover, Janis & Wang, Lei (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science [online]. *Biomedical Digital Libraries*, vol. 3 [dostęp: 11.06.2011]. Dostępny w World Wide Web: <doi:10.1186/1742-5581-3-7>.

Chen, Kuang-hua (2004). The construction of the Taiwan Humanities Citation Index. *Online Information Review,* vol. 28, pp. 410-419. Dostępny w World Wide Web: <doi: 10.1108/14684520410570535>.

Derfert-Wolf, Lidia; Garczyńska, Maria; Matuszewski, Szymon; Rychlewska, Maria (2005). Projekt rejestrowania cytowań w artykułach indeksowanych w „Bazie danych o zawartości polskich czasopism technicznych" BazTech: koncepcja ogólna, 2005 [online]. *E-LIS* [dostęp: 21.05.2011]. Dostępny w World Wide Web: <http://eprints.rclis.org/handle/10760/7260>.

Di Donato, Francesca (2004). Verso uno „European Citation Index for the Humanities": Che cosa possono fare i ricercatori per la comunicazione scientifica [online]. *E-LIS*. [dostęp: 12.05.2011]. Dostępny w World Wide Web: <http://eprints.rclis.org/handle/10760/5629>.

Drabek Aneta; Waga Małgorzata (2009). Możliwości wykorzystania polskich baz danych w ocenie parametrycznej jednostek naukowych. *Sprawy Nauki*, nr 4, s. 27-30.

Drabek Aneta; Tomaszczyk Jacek (2008). Czasopismo „Praktyka i Teoria Informacji Naukowej i Technicznej" w świetle danych bazy CYTBIN. W: *Zarządzanie informacją w nauce*. Katowice: Wydaw. UŚ, s. 365-375.

Garfield, Eugene (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science (New York)*, vol. 122, no. 3159, pp. 108-111.

Garfield, Eugene (1979). *Citation indexing – its theory and application in science, technology, and humanities*. New York: Wiley.

Garfield, Eugene (2004). Historiographic Mapping of Knowledge Domains Literature. *Journal of Information Science,* vol. 30, no. 2, pp. 119-145. Dostępny w World Wide Web: <doi: 10.1177/0165551504042802>.

Giles, C.Lee; Bollacker, Kurt D.; Lawrence, Steve (1998). CiteSeer: An Automatic Citation Indexing System. [In.:] *Digital Libraries 98 – Third ACM Conference on Digital Libraries*, New York: ACM Press, pp. 89-98. Dostępny w World Wide Web: <10.1145/276675.276685>.

Giri, Rabishankar; Das, Anup Kumar (2011). Indian Citation Index: a new web platform for measuring performance of Indian research periodicals. *Library Hi Tech News*, vol. 28, no. 3, pp. 33-35. Dostępny w World Wide Web: <doi: 10.1108/07419051111145154>.

Gong, Fang; Xu, Juan, Fu, Jian, Deng, Sanhong, Bai, Yun (2007). Influences of Chinese educational journals: research based on 2000-2004 CSSCI. *Frontiers of Education in China,* vol. 2, no. 4, pp. 545-567. Dostępny w World Wide Web: <doi: 10.1007/s11516-007-0041-8>.

Hua, Weina (2001). The development of the Chinese Social Sciences Citation Index. *The Indexer*, vol. 22, no. 3, pp. 128-129.

Kolasa, Władysław Marek; Jarowiecki, Jerzy (2005). *Polska bibliografia prasoznawcza 1996-2001*. Kraków: Polska Akademia Nauk.

Kolasa, Władysław Marek (2009). *Uwagi metodologiczne do tworzenia Indeksu Cytowań Historiografii Mediów Polskich.* Wersja 1.7. Kraków: IINiB UP [online]. Scribd [dostęp: 11.06.2011]. Dostępny w World Wide Web: <http://www.scribd.com/doc/11368274>.

Kolasa, Władysław Marek (2011a). Analiza cytowań w naukach historycznych: wybrane problemy i prawidłowości [referat wygłoszony na konferencji „Nauka o informacji (informacja naukowa) w okresie zmian", Warszawa, Uniwersytet Warszawski 4-5.04.2011].

Kolasa, Władysław Marek (2011b). Specific character of citations in historiography (using the example of Polish history). *Scientometrics* [online first 13.11.2011]. Dostępny w World Wide Web: <doi: 10.1007/s11192-011-0553-0>.

Konieczna, Danuta (2002). Bibliometryczna analiza publikacji cytowanych w czasopiśmie „Litteraria" w latach 1969-1999, *Zagadnienia Naukoznawstwa*, z. 1/2, s. 137-145.

Kozłowski, Jan (1994). *Miejsce nauki polskiej w świecie.* Warszawa: Komitet Badań Naukowych.

Lawrence, Steve; Giles C. Lee; Bollacker, Kurt (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, vol. 32, no. 6, pp. 67-71. Dostępny w World Wide Web: <doi: 10.1109/2.769447>.

Lopez, Patrice. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Lecture Notes in Computer Science,* vol. 5714, pp. 473-474. Dostępny w World Wide Web: <doi: 10.1007/978-3-642-04346-8>.

Marshakova-Shaikevich, Irina V. (1996). *Bibliometryczna analiza współczesnej nauki*. Katowice: Wydaw. UŚ.

Marshakova-Shaikevich, Irina V. (2001). Nauka polska u progu XXI wieku: analiza bibliometryczna dorobku Polski i państw Unii Europejskiej, *Roczniki Biblioteczne,* R. 45, s. 147-165.

Marshakova-Shaikevich, Irina V. (2009). *Badania ilościowe nauki: podejście bibliometryczne i webometryczne*. Poznań: Uniwersytet im. A. Mickiewicza.

Nowak, Piotr (2000). *Wybrane problemy efektywności polskich czasopism naukowych z dziedziny humanistyki*. Poznań: Motivex.

Nowak, Piotr (2004). Piśmiennictwo z zakresu nauk społecznych i humanistycznych przedmiotem oceny i analiz metodami bibliometrycznymi: możliwości i ograniczenia. *Roczniki Naukowe PWSZ im. Komeńskiego,* nr 2, ser. A, *Miscellanea*, t. 2, s. 5-18.

Nowak, Piotr (2008). *Bibliometria, webometria: podstawy, wybrane zastosowania*. Poznań: Wydaw. Naukowe Uniwersytetu im. Adama Mickiewicza.

Nwagwu, Williams E. (2010). Cybernating the academe: Centralized scholarly ranking and visibility of scholars in the developing world. *Journal of Information Science*, vol. 36, no. 2, pp. 228-241. Dostępny w World Wide Web: <doi:10.1177/0165551509358482>.

Osiewalska, Anna (2008). Analiza cytowań z wybranych polskojęzycznych czasopism ekonomicznych. [W:] *Zarządzanie informacją w nauce.* Katowice: Wydaw. UŚ., s. 244-256.

Price, Derek John de Sola (1967). *Mała nauka – wielka Nauka*. Warszawa: PWN.

Stefaniak, Barbara, Swoboda, Izabela (2004). Polskie indeksy cytowań – potrzeba tworzenia, dotychczasowe doświadczenia. W: *Piąta Ogólnokrajowa Narada Bibliografów*, Warszawa: Biblioteka Narodowa, s. 244-254.

Stern, Madeleine (1983). Characteristics of the Literature of Literary Scholarship, *College & Research Libraries*, vol. 44, no. 4, pp. 199-209.

Torres-Salinas, Daniel; López-Cózar, Emilio; Jiménez-Contreras, Evaristo (2009). Redes de citación de las revistas españolas de Ciencias Sociales 1994-2006. *Revista Espanola de Documentacion Cientifica*, vol. 32, no. 2, pp. 34-50. Dostępny w World Wide Web: <doi:10.3989/redc.2009.2.686>.

Waga, Małgorzata; Drabek, Aneta (2002). Arton – baza cytowań polskiej literatury humanistycznej (stan prac nad bazą). *Zagadnienia Naukoznawstwa*, R. 38, z. 1/2, s. 147-151.

Webster, Berenika M. (2000). Socjologia polska w świetle Social Sciences Citation Index i Indeksu Cytowań Socjologii Polskiej: analiza porównawcza za lata 1981-1995. *Zagadnienia Naukoznawstwa*, R. 36, z. 2/3, s. 391-417.

Webster, Berenika M. (2001). O potrzebie tworzenia lokalnych indeksów cytowań dla analizy nauk społecznych) ze szczególnym uwzględnieniem socjologii) [online]. *Elektroniczny Biuletyn Informacyjny Bibliotekarzy EBIB*, nr 11 (29) [dostęp: 12.05.2011]. Dostępny w World Wide Web: <http://ebib.oss.wroc.pl/2001/29/bwebster.html>.
Wincławska Berenika M., Wincławski Włodzimierz (1995). Indeks cytowań socjologii polskiej: założenia ideowe i omówienie pierwszych wyników. *Zagadnienia Naukoznawstwa*, R. 31, z. 3/4, s. 243-246.

WŁADYSŁAW MAREK KOLASA
The Institute of Information and Library Studies
Pedagogical University in Cracow
e-mail: wmkolasa@gmail.com

RETROSPECTIVE CITATION INDEX IN HUMANITIES
Concept, method, applications

KEYWORDS: Citation index. Databases. Designing. Methodology. Humanities

ABSTRACT: **Objective** – The method of obtaining a citation index through the transformation of a bibliographical database is discussed. Subsequent stages of database construction are presented and evaluated critically. The following phases are described: systems design, method of database construction, citations construction, conversion, use and maintenance. **Research method** – The article is based on the author's experience gained during the process of constructing the Citation Index for Polish Media Historiography (ICHMP). It includes bibliometric indicators taken from ICHMP and the literature of the field. **Results** – ICHMP index contains 24627 documents interrelated with 63811 citations. The most outstanding advantage of the method in question is its high effectiveness. The database indicators are comparable to those of ISI Index, for instance the citation impact equals 6,7, maximum citations per one publication equal 415 and the winning author has 1075 citations. **Conclusions** – The concept discussed (conventionally named a retrospective index) may be used with success in other humanities fields. Main benefits of the solution involve low cost, high effectiveness and vast computational capabilities.

Artykuł wpłynął do Redakcji 12 lipca 2011 r.