Bibliograficzne bazy danych i ich rola w rozwoju nauki
II Konferencja naukowa Konsorcjum BazTech
Poznań, 17-19 kwietnia 2013

Wojciech Fenrich
Aleksander Nowiński
Katarzyna Zamłyńska
Wojtek Sylwestrzak
Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego
Uniwersytetu Warszawskiego

POL-index — Polska Baza Cytowań

Wojciech Fenrich - absolwent socjologii i filozofii, ukończył studia doktoranckie w Instytucie Socjologii Uniwersytetu Warszawskiego. Od 2011 r. pracuje w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego. Analityk systemu POL-on.

Aleksander Nowiński - kierownik działu rozwoju oprogramowania w Centrum Otwartej Nauki, w ICM Uniwersytetu Warszawskiego. Od lat zajmuje się budowaniem i rozwojem systemów informatycznych dla bibliograficznych baz danych oraz bibliotek wirtualnych. Był m.in. dyrektorem technicznym projektu Europejskiej Matematycznej Biblioteki Cyfrowej, a także koordynuje rozwój platformy Yadda.

Katarzyna Zamłyńska - pracuje w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego od 2003 r. Zajmuje się bibliotekami cyfrowymi, m.in. Polską Matematyczną Biblioteką Cyfrową. Analityk systemu POL-index.

Wojtek Sylwestrzak - pracuje w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego, gdzie kieruje Centrum Otwartej Nauki CeON. Zaangażowany jest w promocję otwartego modelu nauki oraz uczestniczy w szeregu inicjatyw związanych z rozwojem nowoczesnych metod komunikacji naukowej. Jest aktywnym uczestnikiem takich projektów jak OpenAIRE (Europejska infrastruktura otwartego dostępu do badań OpenAIRE), EuDML (cyfrowa biblioteka matematyczna), SYNAT (krajowa e-infrastruktura nauki i techniki), Wirtualna Biblioteka Nauki czy Polska Bibliografia Naukowa (PBN).

Streszczenie: *System POL-index to tworzona w Centrum Otwartej Nauki ICM UW polska baza cytowań. Potrzeba stworzenia narzędzia tego typu stanowiła jeden z wniosków z przeprowadzonej w roku 2012 ewaluacji czasopism naukowych z tzw. listy B, w szczególności tych publikujących artykuły z zakresu nauk humanistycznych i społecznych. System, którego najważniejsze elementy czekają obecnie na wdrożenie, cechuje duża elastyczność w zakresie sposobów pozyskiwania danych. Ich docelowy*

przepływ w systemie opierać się będzie na współpracy z polskimi bazami bibliograficznymi, co pozwoli na minimalizację zaangażowania przedstawicieli czasopism w proces pozyskiwania danych. Dzięki systemowi POL-index możliwe będzie wyznaczenie Polskiego Współczynnika Wpływu, który stanie się elementem ewaluacji czasopism w roku 2014.
Słowa kluczowe: *POL-index, POL-on, Centrum Otwartej Nauki, analiza cytowań, ocena czasopism naukowych, Polski Współczynnik Wpływu*

Abstract: *POL-index — a Polish citation database — is being developed in the Centre for Open Science, a part of Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw. Its development is a result of the evaluation of humanities and social sciences journals from the so-called "list B", conducted in 2012 by the Polish Ministry of Science and Higher Education. The system, whose main elements are currently being prepared for release, is very flexible in terms of data gathering. Its target workflow will be based on cooperation with Polish bibliographic databases, which will minimize the engagement of representatives of scholarly journals. One of the by-products of POL-index system will be the Polish Impact Coefficient, which will be included in the evaluation of scholarly journals in 2014.*
Keywords: *POL-index, POL-on, Centre for Open Science, citation analysis, evaluation of scholarly journals, Polish Impact Coefficient*

Prezentacja

POL-index — podstawowe informacje

System POL-index to polska baza cytowań powstająca w Centrum Otwartej Nauki Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego (ICM UW). Jest elementem systemu informacji o szkolnictwie wyższym POL-on, współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego i pozostającego własnością Ministerstwa Nauki i Szkolnictwa Wyższego (MNiSW). Potrzeba stworzenia systemu gromadzącego informacje o cytowaniach w obrębie czasopism wymienionych w części B ministerialnego wykazu czasopism naukowych[1], stanowiła jeden z kluczowych wniosków płynących z prac Zespołu specjalistycznego do oceny czasopism naukowych MNiSW w roku 2012. O ile kryteria ewaluacyjne opierające się na już istniejących danych, takich jak PIF (Predicted Impact Factor) wyznaczany na podstawie baz Web of Science dla potrzeb ewaluacji, pozwalają na stosunkowo dobre zróżnicowanie czasopism z zakresu nauk ścisłych, technicznych, medycznych i przyrodniczych, o tyle w przypadku czasopism z zakresu nauk społecznych i (przede wszystkim) humanistycznych kryterium PIF okazuje się nieróżnicujące. Niewielka liczba polskich czasopism z zakresu nauk społecznych i humanistycznych indeksowanych w bazach Web of Science, w połączeniu z bardziej lokalnym charakterem tych obszarów nauki (zarówno pod względem poruszanej w nich tematyki, jak i języka publikacji) i odmienną kulturą cytowania (ilość cytowanej literatury, bardzo istotna rola monografii) sprawia, iż

[1] Komunikat Ministra Nauki i Szkolnictwa Wyższego z dnia 20 grudnia 2012 r. Załącznik: część B wykazu czasopism naukowych. W: *Ministerstwo Nauki i Szkolnictwa Wyższego* [on-line]. Warszawa: MNiSW, 2012 [Dostęp 20.07.2013]. Dostępny w World Wide Web: http://www.bip.nauka.gov.pl/.

niezerowym PIF-em mogą poszczycić się jedynie nieliczne czasopisma humanistyczne (mniej niż 10%) i społeczne (mniej niż 5%)[2] z ministerialnej tzw. listy B.

Głównym celem systemu POL-index jest stworzenie narzędzia pozwalającego na uzyskanie informacji o cytowalności czasopism humanistycznych i społecznych z listy B. System będzie otwarty również na inne czasopisma (zarówno polskie, jak i zagraniczne), których wydawcy są zainteresowani dostarczeniem informacji o publikowanych w nich artykułach. W szczególności dotyczy to czasopism z tzw. listy C, w których cytowane są również czasopisma społeczne i humanistyczne ujęte w części B wykazu czasopism punktowanych MNiSW. Pozwoli to na pozyskanie możliwie dużej ilości danych, które w odniesieniu do znacznej liczby czasopism z listy B nie były jak dotąd w systematyczny sposób gromadzone.

Przepływ danych w systemie POL-index

Największym wyzwaniem dla twórców systemu POL-index było opracowanie sposobu pozyskiwania danych umożliwiających identyfikację cytowań. W połączeniu z wymogiem szybkiej realizacji projektu wymuszało to, by system w tym względzie cechowała możliwie duża elastyczność, pozwalająca na zgromadzenie i ewentualne uzupełnianie danych pochodzących z różnych źródeł. Głównym elementem systemu jest jego baza danych, która może być uzupełniana poprzez interfejs WWW, importer plików w prostym formacie XML oraz import ze współpracujących z systemem POL-index polskich baz bibliograficznych. System POL-index składać się będzie ponadto z interfejsu analitycznego pozwalającego na odszukanie informacji o zidentyfikowanych cytowaniach oraz prezentację podstawowych, popularnych wskaźników (takich jak dwuletnia i pięcioletnia miara cytowalności czy indeks Hirscha) oraz systemu maszynowej identyfikacji cytowań opartego o działający w ICM UW klaster Apache Hadoop. POL-index będąc elementem systemu oceny czasopism naukowych, umożliwi też zgłaszanie uwag dotyczących poprawności zidentyfikowanych cytowań. POL-index korzystać będzie z systemu kont użytkowników i przypisanych do nich ról, który funkcjonuje w obrębie Polskiej Bibliografii Naukowej. Pozwoli to w łatwy i spójny sposób zarządzać uprawnieniami osób wypełniających ankietę czasopisma i wprowadzających dane do systemu POL-index.

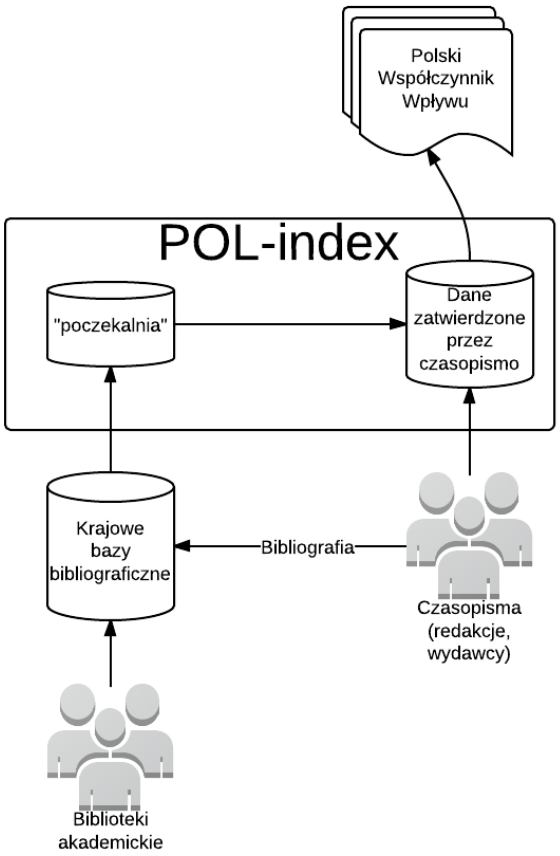
Minimalistyczny wariant przepływu danych w systemie POL-index przewiduje, że całość danych wprowadzana byłaby przez przedstawicieli czasopism w ramach ewaluacji realizowanej corocznie przez Ministerstwo Nauki i Szkolnictwa Wyższego. Przedstawiciele czasopism wnioskujący o ocenę zostaliby w tym wariancie zobligowani

[2] Dane za: WILKIN, J. Ocena parametryczna czasopism naukowych w Polsce — podstawy metodologiczne, znaczenie praktyczne, trudności realizacji i perspektywy. *Forum Akademickie* [on-line]. 14.01.2013 [Dostęp 15.06.2013]. Dostępny w World Wide Web: http://forumakademickie.pl/aktualnosci/2013/1/14/1315/ocena-parametryczna-czasopism-naukowych-w-polsce-podstawy-metodologiczne-znaczenie-praktyczne-trudnosci-realizacji-i-perspektywy/.

(dostarczenie i zatwierdzenie danych w systemie POL-index jako warunek konieczny poddania czasopisma procedurze ewaluacyjnej) lub byliby zachęcani (dostarczenie i zatwierdzenie danych w systemie POL-index jako element umożliwiający uzyskanie wyższej punktacji w ramach procedury ewaluacyjnej) do zdeponowania wymaganych danych. Ich dostarczanie następowałoby równolegle do procesu składania ankiety czasopisma naukowego w Polskiej Bibliografii Naukowej. Po zgromadzeniu danych, w obrębie systemu POL-index następowałaby automatyczna identyfikacja cytowań, a po opublikowaniu jej rezultatów możliwe byłoby zgłaszanie uwag dotyczących jej poprawności i kompletności. Kolejne kroki to powtórna automatyczna identyfikacja cytowań uwzględniająca uwagi zgłaszane przez użytkowników systemu oraz opublikowanie jej ostatecznych wyników.

Optymalny wariant przepływu danych zakłada z kolei, że całość danych byłaby pozyskiwana ze współpracujących z systemem POL-index baz bibliograficznych. Wariant ten wymagałby jednak uprzedniej indeksacji w tych bazach wszystkich czasopism aplikujących o ocenę w ramach części B ministerialnego wykazu czasopism punktowanych. Czasopisma te byłyby dzięki temu zwolnione z obowiązku samodzielnego dostarczania danych do systemu POL-index. Model optymalny zakłada również nawiązanie współpracy z polskimi wydawnictwami naukowymi, co pozwoliłoby uzupełnić bazę systemu o informacje dotyczące książek naukowych, niezmiennie pełniących istotną rolę w naukach humanistycznych i społecznych.

Oprócz wariantu minimalistycznego i optymalnego możliwe do realizacji są również różne kombinacje wariantów pośrednich, w przypadku których część danych byłaby pozyskiwana z baz bibliograficznych współpracujących z POL-index, a część wprowadzana do systemu bezpośrednio przez przedstawicieli czasopism za pomocą interfejsu WWW i/lub importera XML. Dane pochodzące z baz miałyby w tym przypadku charakter pomocniczy i importowane byłyby w postaci tzw. szkiców, tj. rekordów artykułów, które wymagałyby każdorazowo jedynie weryfikacji, zatwierdzenia i ewentualnego uzupełnienia przez przedstawicieli czasopisma (zob. rys. 1). Realizacja tego właśnie wariantu przepływu danych w POL-index w pierwszych latach funkcjonowania systemu jest obecnie (czerwiec 2013 r.) najbardziej prawdopodobna.



Rys. 1. Pośredni wariant przepływu danych w systemie POL-index
Źródło: opracowanie własne.

Polski Współczynnik Wpływu (PWW)

Najistotniejszym rezultatem powstania bazy POL-index w kontekście oceny czasopism naukowych będzie możliwość wyznaczenia Polskiego Współczynnika Wpływu (PWW), który od roku 2014 włączony zostanie w system ewaluacji czasopism z tzw. listy B. Zgodnie z przyjętymi założeniami, PWW powinien być obliczany w sposób możliwie prosty i odwołujący się do już istniejących miar cytowalności. Powinien on jednak uwzględniać specyfikę czasopism humanistycznych i społecznych, w szczególności zaś fakt, iż opublikowane w nich artykuły zaczynają być cytowane po upływie dłuższego okresu, niż ma to miejsce w przypadku publikacji z zakresu nauk ścisłych, technicznych, medycznych i przyrodniczych. Jednocześnie ów okres referencyjny nie może być zbyt długi, tak by PWW mógł odzwierciedlać cytowalność bieżących artykułów naukowych publikowanych w danym czasopiśmie.

Kierując się powyższymi zasadami, Zespół specjalistyczny do oceny czasopism naukowych MNiSW przyjął następujący sposób obliczania Polskiego Współczynnika Wpływu:

$$PWW = \frac{C^{s}_{(s-5,s-1)}}{N^{(s-5,s-1)}} + \frac{C^{s-1}_{(s-6,s-2)}}{N^{(s-6,s-2)}},$$

gdzie:
s — oznacza rok poprzedzający rok złożenia ankiety czasopisma;
$C^{s}_{(s-5,s-1)}$ — oznacza liczbę cytowań w obrębie artykułów opublikowanych w roku s do wszystkich artykułów opublikowanych w czasopiśmie w latach od s-5 do s-1;
$N^{(s-5,s-1)}$ — oznacza liczbę cytowalnych artykułów naukowych opublikowanych w czasopiśmie w latach od s-5 do s-1;
$C^{s-1}_{(s-6,s-2)}$ — oznacza liczbę cytowań w obrębie artykułów opublikowanych w roku s-1 do wszystkich artykułów opublikowanych w czasopiśmie w latach od s-6 do s-2;
$N^{(s-6,s-2)}$ — oznacza liczbę cytowalnych artykułów opublikowanych w czasopiśmie w latach od s-6 do s-2.

Innymi słowy, Polski Współczynnik Wpływu obliczany będzie jako suma dwóch pięcioletnich miar cytowalności wyznaczanych dla każdego z dwóch lat poprzedzających rok oceny czasopism naukowych. Taki kształt PWW określa zarazem niezbędny zakres danych. I tak, w przypadku artykułów opublikowanych w okresie dwóch lat poprzedzających rok oceny czasopism naukowych, w bazie systemu POL-index winny znaleźć się tzw. rekordy pełne, na które składają się informacje o:

- tomie i numerze czasopisma, w którym opublikowano dany artykuł,
- tytule artykułu,
- tytule artykułu w innych językach (o ile tytuł w innym języku zamieszczony był w samym artykule),
- typie artykułu,
- roku wydania,
- języku artykułu,
- imionach i nazwiskach autorów,
- wykazie cytowanej literatury.

Dla artykułów opublikowanych w okresie od trzeciego do siódmego roku poprzedzającego rok oceny czasopism, niezbędne jest pozyskanie tzw. rekordów skróconych, na które składają się informacje o:

- tomie i numerze czasopisma, w którym opublikowano dany artykuł,
- tytule artykułu,

- tytule artykułu w innych językach (o ile tytuł w innym języku zamieszczony był w samym artykule),
- typie artykułu,
- roku wydania,
- języku artykułu,
- imionach i nazwiskach autorów.

Do obu rodzajów rekordów możliwe będzie również wprowadzenie/zaimportowanie informacji o afiliacjach autorów, a do rekordów skróconych również wykazu cytowanej literatury. W drugim przypadku ewentualne cytowania zostaną zidentyfikowane w obrębie systemu, nie będą one jednak uwzględniane przy obliczaniu PWW. Jeśli dany artykuł nie posiada wykazu cytowanej literatury w formie bibliografii załącznikowej, a informacja o cytowanych pozycjach zawarta jest w nim wyłącznie w przypisach dolnych lub końcowych, przedstawiciel czasopisma deponujący dane w POL-index zobowiązany będzie do opracowania takiej bibliografii na podstawie przypisów. Artykuły naukowe, publikowane w tym samym czasopiśmie równolegle w kilku wersjach językowych, będą traktowane jak jeden artykuł. Ze względu na różnice w metadanych poszczególnych wersji językowych (tytuły w różnych językach, inne numery stron), utrudniające poprawną identyfikację cytowań, konieczne będzie jednak wprowadzenie każdej z nich jako osobnego rekordu. Czasopisma takie nie będą więc poszkodowane za sprawą sztucznie zawyżonej wartości mianownika składników PWW. Należy przy tym zaznaczyć, że konieczność pozyskania danych sprzed aż siedmiu lat zachodzić będzie jedynie w pierwszym roku funkcjonowania systemu. W kolejnych latach baza POL-index będzie uzupełniana jedynie o pełne rekordy artykułów z poprzedniego roku.

Teraźniejszość i przyszłość

Obecnie (czerwiec 2013 r.) przygotowuje się uruchomienia takich elementów systemu POL-index, jak interfejs WWW pozwalający na ręczne wprowadzanie danych, możliwość importu w formacie XML oraz możliwość importu z zewnętrznych baz danych. Trwają również końcowe prace związane z budową modułu odpowiedzialnego za automatyczną identyfikację cytowań. Produkcyjna wersja systemu uruchomiona zostanie z odpowiednim wyprzedzeniem, tak by umożliwić spokojne wprowadzenie informacji niezbędnych do określenia cytowalności czasopism w ramach ich ewaluacji w roku 2014.

Do końca 2013 r. ukończone zostaną prace nad częścią analityczną interfejsu WWW. W kolejnych latach prowadzone będą również prace nad poprawą skuteczności algorytmów wykorzystywanych do identyfikacji cytowań w obrębie systemu POL-index. Z myślą o naukach społecznych i humanistycznych planuje się podjęcie starań zmierzających do nawiązania współpracy z polskimi wydawnictwami naukowymi w celu włączenia do bazy systemu rekordów książek naukowych. Wraz z dalszym rozwojem polskich baz bibliograficznych (odpowiednie poszerzenie zakresu metadanych oraz

zwiększenie liczby indeksowanych tytułów naukowych), pozwoli to stopniowo zbliżać się do optymalnego wariantu przepływu danych w systemie POL-index i do niezbędnego minimum ograniczać nakład pracy ponoszony przez przedstawicieli czasopism naukowych.

Bibliografia:

1. Komunikat Ministra Nauki i Szkolnictwa Wyższego z dnia 20 grudnia 2012 r. Załącznik: część B wykazu czasopism naukowych. W: *Ministerstwo Nauki i Szkolnictwa Wyższego* [on-line]. Warszawa: MNiSW, 2012 [Dostęp 20.07.2013]. Dostępny w World Wide Web: http://www.bip.nauka.gov.pl/.
2. WILKIN, J. Ocena parametryczna czasopism naukowych w Polsce — podstawy metodologiczne, znaczenie praktyczne, trudności realizacji i perspektywy. *Forum Akademickie* [on-line]. 14.01.2013 [Dostęp 15.06.2013]. Dostępny w World Wide Web: http://forumakademickie.pl/aktualnosci/2013/1/14/1315/ocena-parametryczna-czasopism-naukowych-w-polsce-podstawy-metodologiczne-znaczenie-praktyczne-trudnosci-realizacji-i-perspektywy/.

Fenrich, W. i in. POL-index — Polska Baza Cytowań. W: Bibliograficzne bazy danych i ich rola w rozwoju nauki. II Konferencja naukowa Konsorcjum BazTech, Poznań, 17-19 kwietnia 2013 [on-line]. Stowarzyszenie EBIB, 2013 [Dostęp: 30.08.2013]. Materiały konferencyjne EBIB, nr 24, Dostępny w World Wide Web: http://open.ebib.pl/ojs/index.php/Mat_konf/article/view/40/. ISBN 978-83-63458-06-5.