nr 6/2014 ISSN 2299-565X
ELEKTRONICZNE CZASOPISMO BIBLIOTEKI GŁÓWNEJ UNIWERSYTETU PEDAGOGICZNEGO W KRAKOWIE

Po czym rozpoznać dobre repozytorium?

Tomasz Lewandowski, Michał Starczewski

Platforma Otwartej Nauki, ICM Uniwersytet Warszawski

Streszczenie

Starannie prowadzone repozytorium to repozytorium dopasowane do całego systemu repozytoryjnego, obejmującego repozytoria instytucjonalne i dziedzinowe oraz repozytoria danych. Wymiana metadanych umożliwia powstawanie agregatorów. Ważna jest nie tylko dostępność treści w Internecie, ale również ich widoczność. Repozytoria mogą poprawiać widoczność zdeponowanych w nich treści uwzględniając wymogi wyszukiwarek internetowych, które dla wielu użytkowników są podstawowym sposobem pozyskiwania informacji naukowej. W artykule przedstawiono generalne zasady poprawiania widoczności w Internecie, a także kilka szczegółowych wskazówek.

Słowa kluczowe

otwarte repozytorium, widoczność, wyszukiwarka, metadane

Na Uniwersytecie Humboldta w Berlinie powstał zespół opracowujący kolejny ranking otwartych repozytoriów. Co wyróżnia go od pozostałych, w tym najbardziej rozpoznawalnego rankingu OpenDOAR? Jego twórcy podkreślają, że ich wyjątkowość polega na skoncentrowaniu się na jakości i staranności prowadzenia repozytorium, zamiast na liczbie zdeponowanych obiektów i spozycjonowaniu w wyszukiwarkach internetowych [1]. Pierwsza edycja rankingu obejmuje wyłącznie repozytoria niemieckie, ale jego twórcy zapowiadają rozszerzenie rankingu również na inne kraje.

Upowszechnienie się niemieckiego rankingu przyczyni się do oceniania repozytoriów według kryteriów odzwierciedlających spełnianie istotnych funkcji tych narzędzi, a nie tylko poprzez proste porównanie wielkości zgromadzonych zasobów.

Przyzwyczajenia naukowców

Narzędzia wspierające badania i komunikację naukową powinny być dostosowane do panujących wśród naukowców zwyczajów. Nawet jeśli stosowanie ich wiąże się z wykształceniem nowych nawyków i koniecznością zdobycia nowych umiejętności, powinno się nawiązywać do tego, co już jest. Próby narzucenia rewolucyjnych zmian raczej skończyłyby się niepowodzeniem i zrażeniem uczonych do nowych narzędzi.

Naukowcy jako grupa wykazują silną tendencję do zachowań konserwatywnych. Również zachowania związane z deponowaniem artykułów naukowych w repozytoriach są mocno związane z tradycją panującą w danej dziedzinie. W społeczności fizyków deponowanie preprintów wszystkich artykułów w arXiv nikogo nie dziwi, a wręcz jest oczekiwane. W innych dyscyplinach takich zwyczajów nie ma. Podobnie, proporcje między liczbą artykułów deponowanych w repozytoriach instytucjonalnych i dziedzinowych wydają się zależeć w głównej mierze od badanej dziedziny nauki [2]. O przyzwyczajeniach naukowców trzeba pamiętać, gdy myśli się o repozytoriach instytucjonalnych i dziedzinowych, ponieważ z punktu widzenia praktyk komunikacyjnych naukowców repozytoria mogą być dla siebie konkurencją lub wspierać się wzajemnie. Zakłada się nieraz, że deponowanie w jednym repozytorium wykształca nawyk deponowania, który sprawia, że naukowiec deponuje również w innych repozytoriach poza macierzystym. Badania przeprowadzone wśród fizyków przez Jingfeng Xia wskazują jednak, że może pojawić się także inna zależność: naukowiec deponujący w jednym repozytorium, niezależnie od tego, jakiego jest ono typu, nie będzie już skłonny deponować w innym [3].

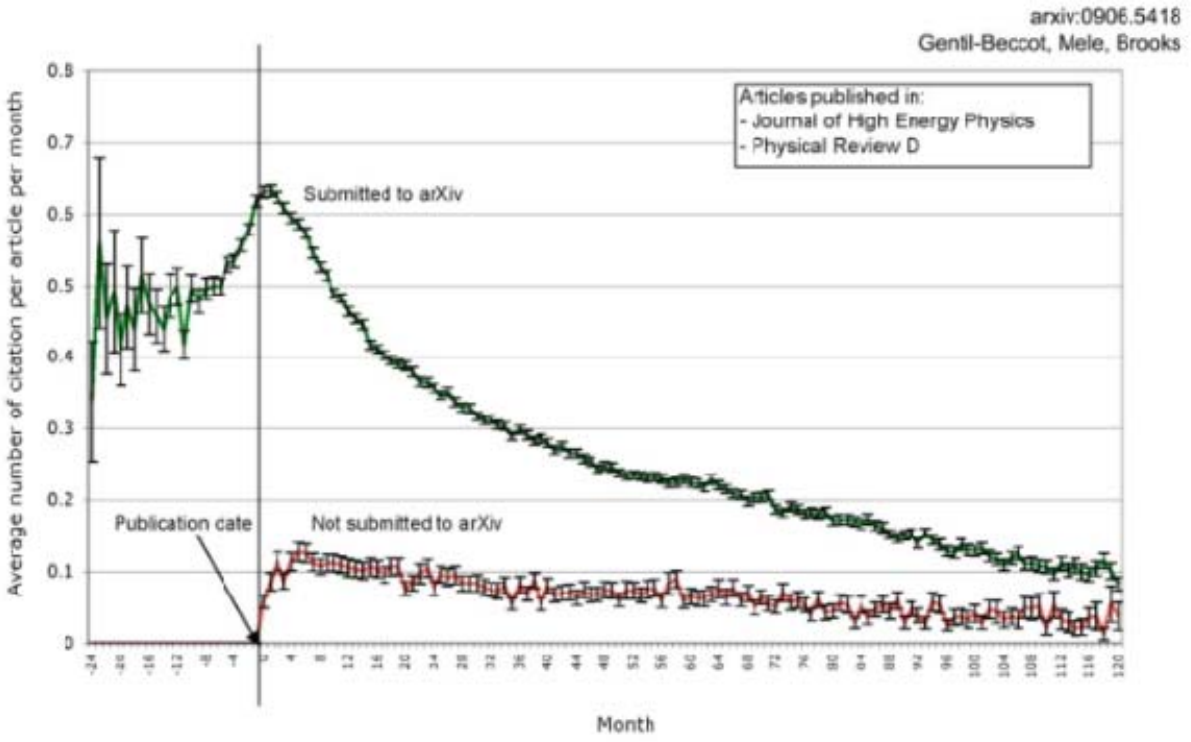
System repozytoryjny

Repozytorium, czyli narzędzie informatyczne służące do deponowania, gromadzenia i udostępniania przede wszystkim bieżącego dorobku naukowego [4], należy wyraźnie odróżnić od bibliotek cyfrowych (będących elementem systemu obiegu treści dziedzictwa kultury) oraz baz i platform udostępniających bieżące treści naukowe (prowadzone przez redaktorów w sposób systematyczny). W repozytoriach deponują swoje prace sami naukowcy. Zadaniem redaktorów jest nie tyle zapewnienie kompletności materiałów, co troska o jakość metadanych i dobrą widoczność w Internecie. Starannie prowadzone repozytorium to repozytorium, które jest skutecznie włączone do sieci międzynarodowej infrastruktury repozytoryjnej. Spełnia standardy, dzięki którym zdeponowane materiały są łatwe do odnalezienia.

System repozytoryjny składa się z kilku kategorii narzędzi.

Repozytoria instytucjonalne są zakładane przez instytucje i służą przede wszystkim do gromadzenia, udostępniania i promowania publikacji, których autorzy są pracownikami danej instytucji. Nie jest tu istotna dziedzina badań, choć z reguły struktura instytucji jest odzwierciedlona w strukturze repozytorium, np. każdy wydział ma swoją kolekcję. Repozytoria instytucjonalne pomagają zarządzać instytucją, dostarczając informacji o dorobku naukowym. Umożliwiają prowadzenie szczegółowych statystyk.

Repozytoria dziedzinowe są dedykowane wszystkim naukowcom zajmującym się badaniami w określonej dyscyplinie, bez względu na afiliację. Mniej istotne jest w ich przypadku wspieranie zarządzania uczelnią, za to mogą organizować społeczność naukowców zajmujących się danym tematem. Sztandarowym przykładem udanego repozytorium dziedzinowego jest arXiv, które dla wielu fizyków jest podstawowym źródłem informacji o nowych publikacjach. Artykuły, które oprócz tradycyjnej publikacji zostały zdeponowane w tym archiwum jako preprinty są statystycznie rzecz biorąc o wiele częściej cytowane niż artykuły, które ukazały się wyłącznie w czasopiśmie (patrz Rys. 1) [5]. Repozytoria dziedzinowe mogą mieć wąskie specjalizacje i naszym zdaniem prawdopodobny jest rozwój wyspecjalizowanych repozytoriów dziedzinowych, będących ośrodkami współpracy poszczególnych środowisk uczonych.



Rys. 1. Średnia liczba cytowań wraz z upływem czasu przed i po publikacji w czasopiśmie - dla artykułów z dziedziny fizyki wysokich energii - z podziałem na te opublikowane w arXiv i w czasopiśmie oraz te opublikowane wyłącznie w czasopiśmie. Źródło: Gentil-Beccot A., Mele S., Brooks T. C., *Citing and reading behaviours in high-energy physics : how a community stopped worrying about journals and learned to love repositories*, **"Scientometrics", [online], 2009, Vol. 84, nr 2, s. 345-355 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://arxiv.org/ftp/arxiv/papers/0906/0906.5418.pdf.**

W pewnym sensie takim bardzo wyspecjalizowanym repozytorium jest *repozytorium projektowe* (np. Repozytorium ECNIS) [6]. Duży projekt, realizowany przez szereg instytucji, może posługiwać się repozytorium w celu usprawnienia komunikacji i zbiorczego przedstawienia efektów projektu. Wszystkie publikacje (a także dane i inne materiały) wytworzone w ramach projektu mogą zostać pokazane jako spójny zbiór, zamiast rozproszenia w wielu kanałach dystrybucji treści naukowych.

Repozytoria danych to szczególna kategoria repozytoriów. Jest ich wciąż o wiele mniej niż repozytoriów publikacji. Udostępnianie danych badawczych przed epoką Internetu napotykało na bariery techniczne. Zarówno instytucje wspierające badania, jak i sami naukowcy muszą wciąż się uczyć, jak najlepiej korzystać z tej możliwości. Udostępnienie danych wymaga dodatkowej pracy, która nie jest uwzględniana w ocenie dorobku naukowego. Repozytoria danych będą się mimo wszystko rozwijały, czemu powinno pomóc wprowadzanie polityk instytucjonalnych, obligujących beneficjentów programów finansujących badania nie tylko do otwartego udostępniania publikacji, ale również surowych danych.

Choć póki co mówi się o repozytoriach danych jako takich, to spodziewamy się, że wraz ze wzrostem ich liczby zacznie się dzielić je na instytucjonalne i dziedzinowe, tak jak ma to miejsce w przypadku repozytoriów treści.

Dlaczego nie wystarczą repozytoria służące do deponowania zarówno publikacji, jak i danych? Obecnie część repozytoriów treści udostępnia przecież również dane. Oddzielenie treści od danych ma jednak głębokie uzasadnienie. Metadane publikacji różnią się znacząco od metadanych danych badawczych, tym bardziej, że dane badawcze mogą przyjmować bardzo różne formy. Często są to całe kolekcje danych. Ważną cechą odróżniającą treść od danych jest wielkość plików. Publikacje są niewielkie, podczas gdy dane potrafią przyjmować objętość wielu gigabajtów.

Patrząc na repozytoria jako na elementy systemu repozytoryjnego należy unikać perspektywy wyścigu, w którym kryterium jest to, kto zgromadzi więcej obiektów. O wiele wartościowsza jest perspektywa współpracy. Ewentualna rywalizacja powinna dotyczyć widoczności zdeponowanych materiałów.

Samo założenie repozytorium nie wystarcza do tego, by stało się ono w sposób skuteczny elementem systemu repozytoryjnego i spełniało efektywnie swoje funkcje, inne niż tylko gromadzenie i archiwizacja. Autorzy raportu *Otwarta nauka w Polsce 2014. Diagnoza* zapytali redaktorów polskich repozytoriów o wagę przykładaną przez nich do poszczególnych funkcji. Udostępnianie dorobku naukowego oraz promocja instytucji znalazły się w grupie uznawanej za bardzo ważne lub ważne [7]. Obie te funkcje mogą być dobrze realizowane jedynie przy trosce o widoczność repozytorium w sieci.

Agregatory repozytoriów są ważnym elementem systemu repozytoryjnego. Mimo, że nie gromadzą żadnych pełnych tekstów, dzięki pobieraniu metadanych z wielu repozytoriów stanowią cenne źródło informacji. Agregatory mogą stanowić rodzaj dynamicznie rozwijających się tematycznych bibliografii. Jeśli repozytoria mają dobrze przygotowane metadane, to zbudowanie takiego agregatora jest proste i tanie.

Jak zauważa Marcin Werla, współpraca repozytorium z agregatorem oprócz eksportu odpowiednio przystosowanych metadanych wymaga najczęściej także formalnych porozumień [8]: „Współpraca z agregatorami danych najczęściej oznacza konieczność spełnienia formalnych i technicznych wymogów stawianych przez te agregatory. W związku z tym często niezbędne jest publiczne udostępnienie danych na określonych zasadach oraz dostosowanie danych do schematu i wymogów agregatora. Wymagać to może mapowania oraz wzbogacania danych."

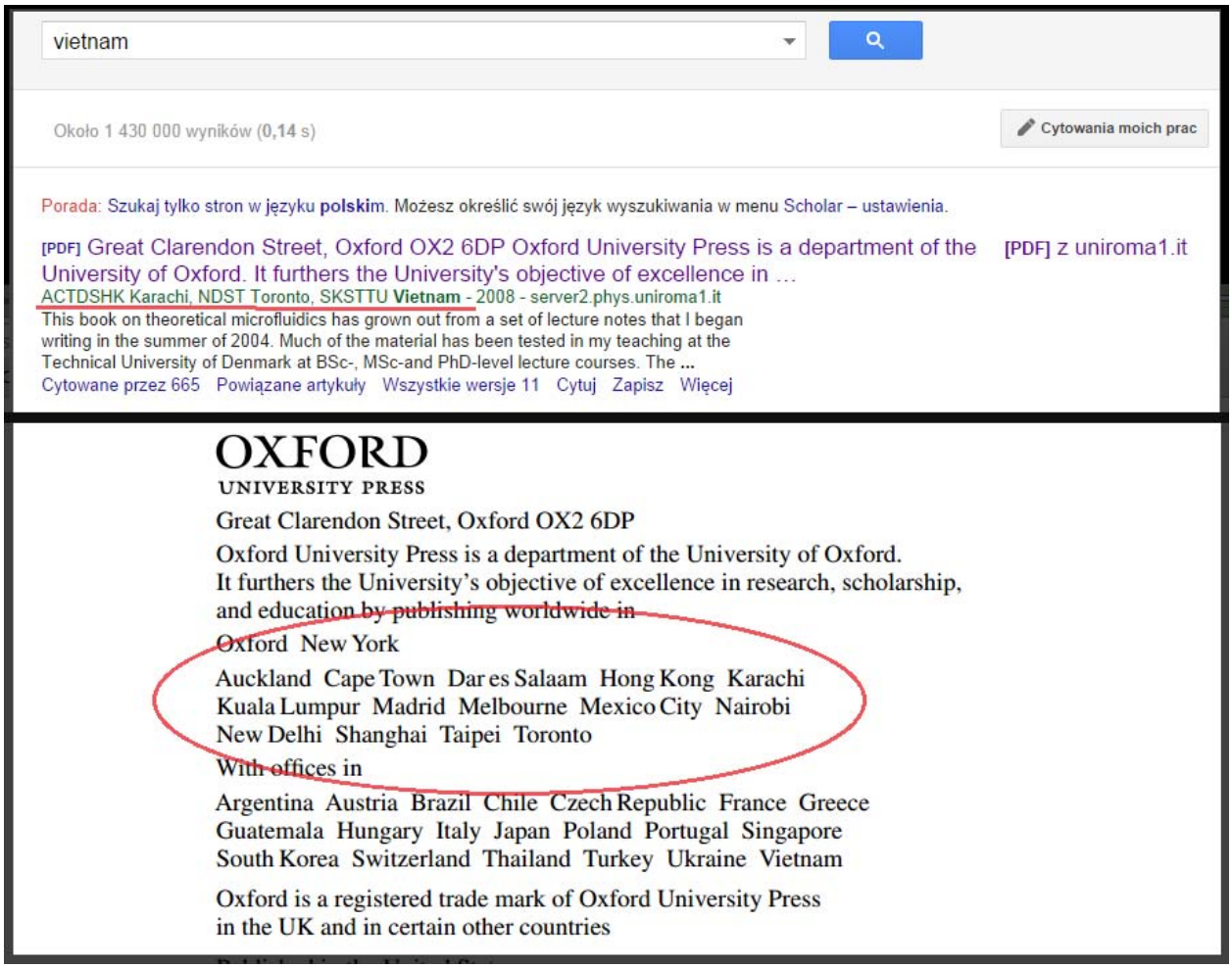
To, co sprawia, że repozytoria nie są niezależnymi magazynami treści naukowych, ale spójnym systemem komunikacji naukowej, to *metadane* i *protokoły* do ich wymiany. Utrzymanie wysokich standardów w zakresie metadanych jest fundamentem systemu repozytoryjnego. To one pozwalają na tworzenie agregatorów i zapewniają elastyczność w ich planowaniu. Treści, które zostają zdeponowane w repozytorium instytucjonalnym mogą być wykorzystywane w zbiorach tematycznych lub dziedzinowych.

W opublikowanym 30 września 2014 roku raporcie Grupy Zadaniowej ds. Metadanych (Metadata Task Force), istniejącej w ramach Konsorcjum World Wide Web czytamy, że wydawcy treści naukowych traktują problem metadanych tak, jakby był już rozwiązany. W wywiadach często dają wyraz przekonaniu, że system spójnych i wzajemnie zgodnych standardów zyskał zasięg, który można uznać za uniwersalny, a większość przeszłych trudności została rozwiązana [9].

Choć nie sposób temu stanowisku odmówić dużej dozy słuszności, trzeba dodać, że nie wszystkie problemy związane z metadanymi zostały rozwiązane. Każdy też, kto w ramach pracy w bazie bibliograficznej czy jako autor tekstu konstruował metadane dla artykułu naukowego wie, że wysoka jakość metadanych nie przychodzi sama, bez kosztów i włożonej pracy.

Świadectwem tego, że dyskusja wokół metadanych nadal żywo się toczy jest np. niedawny tekst Erica van de Velde na jego blogu „SciTechSociety - *The Metadata Bubble* [Van de Velde 2014]. Autor w artykule argumentuje, że aktualny system wprowadzania metadanych w zgodności z wieloma istniejącymi standardami jest nie tylko drogi w utrzymaniu, ale i nie jest w stanie nadążyć za zmieniającą się technologią. Co więcej, już teraz nie wykorzystujemy wszystkich jego możliwości - najpowszechniej eksploatowanym formatem spośród zgodnych z OAI-PMH jest wieloznaczny i niepełny Dublin Core. Zapewniający wymianę i powtórne użycie OAI-ORE jest wykorzystywany zbyt rzadko. Zamiast polegać na wprowadzanych ręcznie danych, powinniśmy zbudować system ich automatycznej ekstrakcji z gotowych artykułów [10].

Niemal automatyczną odpowiedzią każdego, kto miał styczność z metadanymi dostarczanymi przez Google Scholar (a jest to, choć Van de Velde nie wspomina o tym explicite, najpowszechniej dziś chyba wykorzystywany przykład automatycznie ekstrahowanych metadanych) jest wytykanie błędów, jakie powstają podczas ekstrakcji tych danych (przykład: Rys. 2). Google Scholar znany jest z wielu takich błędów [11], wiadomym faktem jest też to, że metadane w nim indeksowane nie są w żaden sposób sprawdzane przez zespół Google. Na to z kolei zwolennicy automatycznie konstruowanych metadanych mogliby odpowiedzieć, że standaryzacja artykułów naukowych (wygląd pierwszej strony, jeden format cytowań bibliograficznych itp.) znacznie podwyższyłaby skuteczność ekstraktorów. Ponadto, mimo głośnych błędów, Google Scholar dotychczas wciąż rósł i zyskiwał na znaczeniu.



Rys. 2. Jeden z bardziej znanych błędów w bazie danych Google Scholar. W dolnej części - fragment źródłowego pliku PDF. Oznaczone czerwoną elipsą nazwy krajów zostały rozpoznane jako imiona i nazwiska autorów. W rezultacie fikcyjne osoby ACTDSHK Karachi, NDST Toronto i SKSTTU Vietnam zostały zaindeksowane jako autorzy. Jak widać w górnej części ilustracji, tytuł pozycji również został błędnie sparsowany. Aktualnie [02-10-2014], aby zobaczyć ten błąd wystarczy wpisać „vietnam w pasku wyszukiwań Google Scholar (http://scholar.google.pl/scholar?hl=pl&q=vietnam). Źródło: opracowanie własne.

Dyskusja na ten temat zapewne będzie się toczyła jeszcze przez jakiś czas. Jest ona częścią tego samego sporu, jaki trwa między zwolennikami samokierujących się samochodów i algorytmów tworzących muzykę a obrońcami ludzkiego nadzoru nad zawodnym oprogramowaniem. Jedno jest niemal pewne: wpisywane ręcznie czy automatycznie ekstrahowane, w każdym wypadku metadane są podstawą współpracy między repozytoriami.

Zasoby naukowe w głębokim Internecie

Jeszcze dekadę temu [12] panowało powszechne przekonanie, że zasoby naukowe należą do tzw. sieci głębokiej (ang. *Deep Web*), nie indeksowanej przez popularne wyszukiwarki internetowe. Sieć głęboka jest pojęciem nieostrym. Zakwalifikowanie treści do głębokiego Internetu zależy od listy „popularnych wyszukiwarek", co często sprowadza się w praktyce do uznania, że to, co nie jest indeksowane przez wyszukiwarkę Google, znajduje się w głębokiej sieci. Jest to, oczywiście, zbyt uproszczony obraz. W ostatnich latach dwa procesy sprawiają, że powoli zmienia się ten stan rzeczy. Po pierwsze, wraz z rozwojem technologii najpopularniejsze wyszukiwarki na czele z wciąż podnoszącym konkurencji poprzeczkę Google, zaczynają pokrywać obszary Internetu znacznie przekraczające to, co wcześniej wydawało się niemożliwe. Po drugie, pojawiły się wyspecjalizowane wyszukiwarki akademickie - na czele z Google Scholar (w 2004 roku) i Microsoft Academic Search (2009) [13]. W przeciwieństwie do baz danych (takich jak np. Highwire), które wcześniej pełniły rolę swoistych przewodników po zasobach naukowych, wyszukiwarki te uzupełniają swoje bazy danych w sposób automatyczny, z pomocą algorytmów „pełzających po sieci w sposób analogiczny do tych znanych z wyszukiwarek o przeznaczeniu ogólnym.

To powolne wynurzanie się elektronicznych zasobów naukowych z głębin sieci sprawiło, że użytkownicy masowo przenieśli się z baz takich jak Highwire do wyszukiwarek akademickich [14]. Zmiana zachowań użytkowników szukających treści naukowych sprawia, że dla osób odpowiedzialnych za zasoby naukowe, w tym repozytoria, ogromnej wagi nabiera widoczność w wyszukiwarkach internetowych, dostosowywanie serwisów internetowych do wymagań Google i wszystkie zagadnienia związane z tzw. SEO (*search engine optimization*, czyli pozycjonowanie stron), które dotąd były domeną serwisów komercyjnych i blogerów. Skoro zdecydowana większość użytkowników zaczyna i kończy wyszukiwanie treści naukowych w popularnej wyszukiwarce [15] nagle okazało się, że dobrze prowadzona strona domowa może być bardziej widocznym źródłem treści niż niedostosowane pod względem SEO profesjonalne repozytorium [16].

Wzrost znaczenia wyszukiwarek internetowych w życiu akademickim sprawił, że na naszych oczach zmienia się sposób, w jaki repozytoria treści naukowych pełnią swoją funkcję. Dawny model ogromnego, samowystarczalnego repozytorium okazuje się nieskuteczny. Organizacje takie, jak OARR (Open Access Repository Ranking – Ranking Otwartych Repozytoriów) podkreślają inny model: niekoniecznie wielkich, lecz dobrze zarządzanych, widocznych w sieci i współpracujących ze sobą repozytoriów.

Warto zwrócić uwagę na fakt, że ranking OARR grupuje razem repozytoria instytucjonalne i dziedzinowe [17]. Podkreśla to, naszym zdaniem, że oba typy repozytoriów są niezbędne do stworzenia wydajnego systemu repozytoryjnego.

Zapatrzenie w największe repozytoria, takie jak PubMed czy arXiv, może odwrócić uwagę od faktu, że małe, ale starannie prowadzone repozytoria potrafią być bardzo dobrze widoczne. Tymczasem oba wielkie repozytoria okazały się zbyt duże, by istnieć jako monolity. arXiv stopniowo wydzielało oddzielne kolekcje. Aktualnie obejmuje pięć „subrepozytoriów" poświęconych innym dyscyplinom niż fizyka, dla której zarezerwowany jest główny adres arxiv.org. Również PubMed jest swoistym klastrem repozytoriów. Pod szyldem PubMed znajduje się kilkanaście wyodrębnionych kolekcji.

Starannie prowadzone repozytorium

Terminem SEO oznacza się szeroki wachlarz rozwiązań z zakresu architektury stron internetowych i zarządzania nimi. Ogólnie rzecz biorąc, wszystko, co może wpłynąć na wzrost oceny strony przez wyszukiwarki internetowe, a przez to - na pozycję, na której dana strona internetowa pojawia się w odpowiedzi na zapytanie użytkownika - będzie przedmiotem analizy SEO. Poniżej wymieniamy kilka najważniejszych czynników, jakie wpływają na ocenę serwisu „w oczach algorytmów oceniających, używanych przez główne wyszukiwarki. W żadnym wypadku nie jest to lista kompletna. Warto jednak zacząć od wyjaśnienia, dlaczego w ogóle wyszukiwarka ocenia strony internetowe, a ogólniej - jak działa.

Wyszukiwarki internetowe to serwisy służące użytkownikom do znajdywania nowych treści w Internecie. Użytkownik wysyła do wyszukiwarki jedno lub więcej słów kluczowych, na podstawie których wyszukiwarka buduje tzw. SERP (*Search Engine Results Page*, czyli po prostu stronę z wynikami z wyszukiwarki), składającą się z linków proponowanych stron wraz z krótkim opisem każdej z nich. Pierwszym celem prowadzenia przez serwisy wyszukiwawcze swoistego „rankingu stron" jest więc intencja wyświetlenia jak najlepszych wyników, będących z najwyższym prawdopodobieństwem użytecznymi dla klienta. Wyszukiwarka oczywiście nie poszukuje linków na bieżąco po każdym zapytaniu, ani nie przeprowadza ich ewaluacji w czasie rzeczywistym - inaczej aktualny poziom prędkości ich działania nie byłby możliwy do otrzymania. Wyszukiwarka obsługuje zapytania na podstawie już gotowej bazy danych. Produkcją tej bazy zajmują się tzw. *crawlery*, po polsku zwane robotami lub pająkami - algorytmy odczytujące zawartość strony internetowej i podążające śladem znajdujących się w jej hipertekście linków do innych stron [18]. Na serwerach danego serwisu w każdym momencie działa wiele instancji różnorakich pająków. Możliwości obliczeniowe tych serwerów nie są oczywiście nieograniczone, wbrew pozorom, jakie mogą powstać przez porównanie z jakimikolwiek uniwersyteckimi serwerami. Celem serwisu jest zindeksowanie jak największej ilości zasobów internetowych (Google ze swoim ogromnym zasięgiem ustanowił tutaj nowe standardy i nie da się łatwo zepchnąć z pozycji lidera, pozostałe wyszukiwarki zaś usiłują mu pod tym względem dorównać, konkurencja jest więc zacięta). Jeśli pająk zapętli się w strukturze danej strony internetowej, zbyt długo musi czekać na odpowiedź jej serwera lub jego algorytmy heurystyczne wskazują, że strona jest mało ciekawa dla użytkownika, nie jest w interesie serwisu kierowanie ograniczonego zasobu obliczeniowego na indeksowanie danej strony [19]. Jest to drugim głównym powodem, dla którego wyszukiwarki internetowe starają się oceniać zasoby - muszą się same orientować, co się opłaca indeksować, a co nie.

Te dwa dominujące czynniki jednocześnie determinują sposób oceny stron przez wyszukiwarki. Podsumowując, na wystawianą przez wyszukiwarkę ocenę strony składają się dwa główne czynniki. Pierwszym z nich jest atrakcyjność dla użytkowników. Tutaj podstawą oceny jest tzw. *PageRank,* algorytm oceniający strony w zależności między innymi od liczby linków do niej prowadzących, z uwzględnieniem analogicznej oceny dostarczających linków stron [20]. Drugim czynnikiem jest atrakcyjność dla serwisu wyszukiwawczego (bezawaryjny serwer, *sitemap*, zgodność ze standardami W3C czyli World Wide Web Consortium - organizacji odpowiedzialnej za ustanawianie standardów pisania i przesyłu stron WWW [21] oraz ogłaszanymi rekomendacjami [22]), który będzie indeksować więcej zasobów strony i będą się one pojawiać wyżej (wcześniej) w wynikach wyszukiwania.

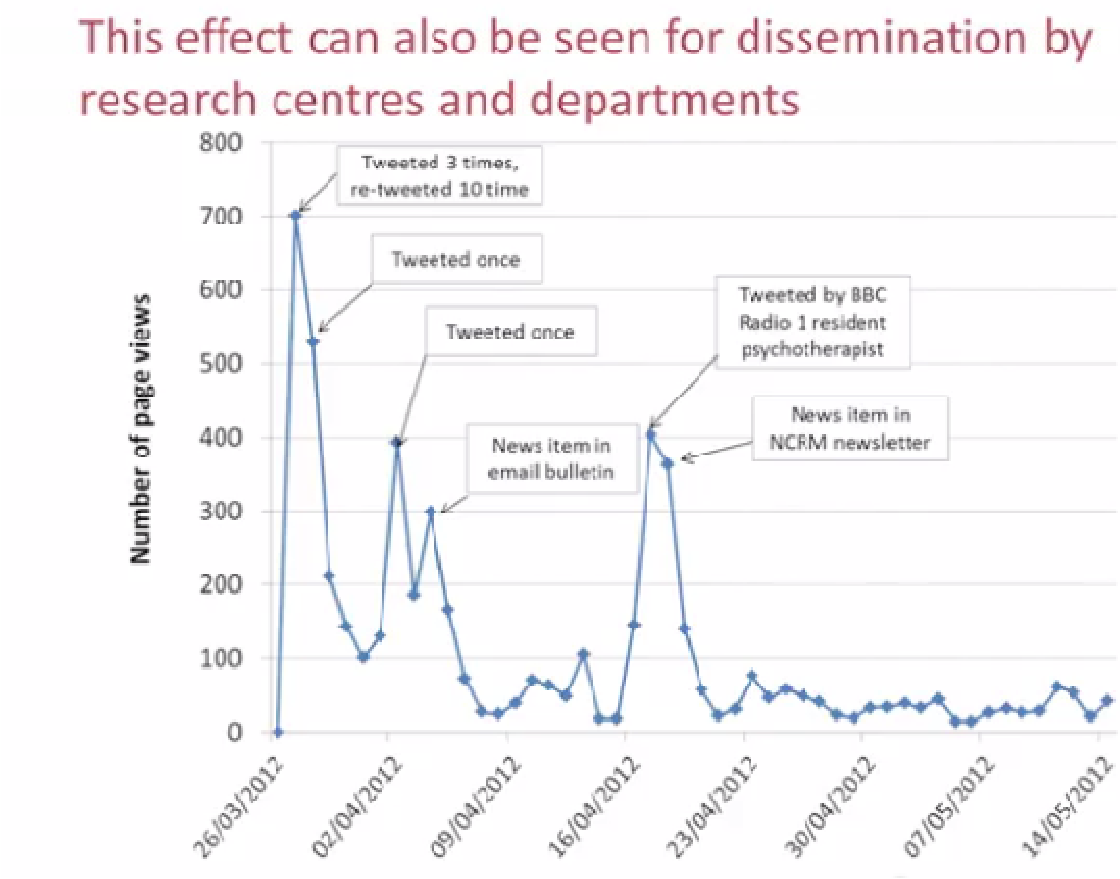
Repozytoria, na szczęście, nie muszą startować w biznesowym „wyścigu szczurów" o cenne pierwsze miejsce w wynikach wyszukiwania. SEO dla repozytoriów ma za to specyficzne problemy. Z punktu widzenia wyszukiwarki internetowej repozytorium jest domeną relatywnie dużą, o zasobach zmieniających się w niewielkim stopniu (ma to wpływ m.in. na ranking Google, preferujący zasoby dynamiczne w stylu *social media*) [23]. Po każdym wywołaniu kolejnego URL z rekordem lub artykułem trzeba czekać na odpowiedź bazy danych. Jeśli nie dołoży się szczególnych starań, żeby „pełzanie" po zasobach repozytoryjnych było dla robota łatwe, najprawdopodobniej zrezygnuje on z indeksowania większości zasobów [24]. Tymczasem w żywotnym interesie repozytoriów jest to, by zaindeksowana została całość zasobów. Zasoby niezindeksowane pozostaną niewidoczne dla większości użytkowników. Trudno jest wytłumaczyć autorom, dlaczego akurat ich artykuł nie pojawił się jeszcze w Google Scholar.

Nie będziemy tu omawiać szczegółowo zmian, jakie trzeba wprowadzić w serwisie internetowym, by stał się przyjazny dla robotów wyszukiwarek (a pośrednio i dla użytkowników). Ten temat poruszany jest w innych artykułach, których kilka propozycji znajduje się w bibliografii.

Najtrudniejszym zadaniem jest zbudowanie zespołu kompetentnego w kwestii SEO, co wymaga dostrzeżenia problemu przez osoby odpowiedzialne za obsługę i finansowanie repozytorium.

Wystarczy przywołać doświadczenia repozytorium instytucjonalnego Uniwersytetu w Utah. Samo przekonwertowanie metadanych na standardy polecane przez Google zwiększyło procent indeksowanych zasobów z 18% do 98% (dla „zwykłego" Google - dla Google Scholar materia była bardziej skomplikowana i samo przekonwertowanie metadanych nie wystarczyło) [25].

Wspomnieliśmy, że jednym z głównych czynników branych pod uwagę podczas obliczania rankingu strony jest liczba linków, które do niej prowadzą. Błędne koło zamyka się gdy do wysokiego rankingu w wyszukiwarkach potrzeba, aby użytkownicy docenili zdeponowane w nich publikacje - jednak aby do nich dotarli, konieczny jest wysoki ranking w wyszukiwarkach. Dostrzegamy tutaj dodatkowo pewną charakterystyczną trudność, z jaką muszą borykać się repozytoria zasobów naukowych. Link do artykułu w repozytorium nie jest tradycyjną formą cytowania publikacji naukowych. Nawet jeśli w repozytorium zostały zdeponowane same artykuły z najwyższej półki, tj. posiadające ogromną liczbę cytowań, nie wpłynie to na sam ranking strony. Dlatego bardzo potrzebne do zaistnienia w przestrzeni internetowej jest otoczenie repozytorium „chmurą" odniesień w mediach społecznościowych, na blogach, na stronach Wikipedii - zaciekawionych użytkowników należy przekierować z miejsc, które już odwiedzają [26]. Korzyści są podwójne. Jak widać z powyższych przykładów, im więcej odniesień, tym więcej użytkowników wie o repozytorium i korzysta z niego. Użytkownicy ci pozostawiają z kolei namacalny ślad w postaci dalszych odniesień do zawartości repozytorium, co dodatnio wpływa na ranking repozytorium w wyszukiwarkach internetowych (co znowu podnosi liczbę zapoznanych z repozytorium użytkowników). Zamiast błędnego koła zaczyna działać sprzężenie zwrotne.



Rys. 3. Wpływ mediów społecznościowych na rozpowszechnienie artykułu naukowego. Źródło: Tinkler J., *Open Access + social media = increased downloads*. **In** *YouTube* **[online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: https://www.youtube.com/watch?v=inYzQABuJ-Y.**

O technicznych aspektach współpracy repozytoriów z mediami społecznościowymi pisze syntetycznie Marcin Werla [27]: „Informacje z baz danych trafiają do [serwisów informacyjnych i społecznościowych] najczęściej poprzez samodzielne kopiowanie lub udostępnianie danych przez użytkowników. W związku z tym ważne jest ułatwienie użytkownikom kopiowania metadanych oraz dostarczanie prostych i trwałych odnośników. Dzięki temu zwiększa się szansa na to, że dane kopiowane z bazy danych będą opatrzone również linkami zwrotnymi przyciągającymi nowych użytkowników. W tym kontekście ważne jest też wsparcie bazy danych dla wykorzystywania zewnętrznych identyfikatorów, takich jak np. DOI."

Zakończenie

Niemiecki ranking repozytoriów proponuje zwracanie uwagi na szereg kryteriów innych niż tylko liczba zgromadzonych obiektów. Ocena repozytoriów sporządzona na podstawie 50 kryteriów podzielonych na 6 kategorii (http://repositoryranking.org/?page_id=660) obejmuje faktycznie ważne zagadnienia, decydujące o tym, czy dane repozytorium spełnia swoje funkcje.

Kryteria te nie mogą być jednak statyczne, z czego niemiecki zespół zdaje sobie w pełni sprawę. Rozwój technologii oraz zmiany wprowadzane w wyszukiwarkach internetowych sprawiają, że kryteria te należy nieustannie krytycznie oceniać i dostosowywać.

Nawet jeśli niemiecki ranking nie znajdzie uznania, nie umniejszy to znaczenia dążenia do jak najlepszej widoczności i oferowania wysokiej jakości usług, dzięki czemu korzystanie z repozytorium będzie pomocne zarówno dla autorów, jak i dla szukających treści naukowych. Współpraca między repozytoriami bazująca na wymianie metadanych pozwoli budować narzędzia dostosowane do konkretnych społeczności badaczy.

Przypisy:

[1] „It is not the size of an open access repository that matters but the quality of the service". Zob. *Metric*. In *2014 Open Access Repository Ranking* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://repositoryranking.org/?page_id=660.

[2] Zob. [Laakso, Björk 2012] - w artykule przeprowadzono m.in. porównanie wyników pięciu badań dotyczących tendencji w deponowaniu publikacji naukowych przez naukowców przeprowadzonych na przestrzeni czterech lat (2009-2012). Różnica wyników między badaniem dotyczącym wyłącznie naukowców z dziedziny inżynierii budownictwa a badaniem dotyczącym naukowców ze wszystkich dziedzin była znaczna zarówno pod względem używania systemów archiwizujących innych niż repozytoria (74% do 26% u inżynierów na korzyść zwykłych stron internetowych, 33% do 67% wśród ogółu naukowców), jak i pod względem proporcji artykułów deponowanych w repozytoriach instytucjonalnych i dziedzinowych (odpowiednio 23% do 3% u inżynierów i 24% do 43% wśród ogółu naukowców). O zwyczaju deponowania przez fizyków w arXiv por. np. [Gentil-Beccot, Mele, Brooks 2009].

[3] [Xia 2008].

[4] [Szprot 2014], s. 30.

[5] Wykres z [Gentil-Beccot, Mele, Brooks 2009] pokazuje jeszcze jeden ciekawy fakt: szczyt cytowania artykułu, którego preprint został zdeponowany w arXiv przypada mniej więcej na czas publikacji - od tego momentu średnia cytowań zaczyna spadać. Oczywiście, te artykuły, które nie zostały zdeponowane w arXiv dopiero od momentu publikacji zaczynają zbierać cytowania.

[6] *ECNIS Repository* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://ecnis.openrepository.com/ecnis/.

[7] [Szprot 2014], s. 36.

[8] [Werla 2013], s. 6.

[9] [Metadata Task Force Report. W3C 2014]: "Scholarly publishers see metadata mainly as a „solved problem": while none would assert that the current situation is perfect, the standards-based consensus in the scholarly publishing world-consisting of nearly universal participation in CrossRef and CCC (the Copyright Clearance Center), the ubiquitous use of the JATS XML model2 for markup and metadata, and the reliance upon the DOI as a persistent, actionable identifier - initially for journal articles but now increasingly for book chapters and components, reference content, conference proceedings, and other publications, as well as the data sets that support research) - has enabled the development a rich ecosystem of services and platforms that has made the Web the primary mode of publication, dissemination, and access for scholarly content. It has also led to the development of other standards - such as ORCID, the Open Researcher and Contributor ID, and FundRef, a system for making public the funders of research - that continually refine the sophistication and utility of metadata in the scholarly publishing world, solving what were previously significant pain points (e.g., disambiguating contributor identities, revealing potential conflicts of interest or reliably documenting the absence of such conflicts)."

[10] Jednym z narzędzi służących do tego celu jest program Cermine, ekstraktor metadanych z plików PDF, powstały w Centrum Otwartej Nauki ICM UW. Zob. *CERMINE* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://cermine.ceon.pl/.

[11] Dość wyczerpującego wyliczenia bardziej strukturalnych błędów dostarcza np. Peter Jacsó w [Jacsó 2006] i [Jacsó 2008].

[12] Zob. [Bergman 2001], zwłaszcza Tabela 2. Także np. implicite [Hatala M. et al. 2004].

[13] [Ortega, Aguillo 2014], s. 1.

[14] Por. np. [Cothran 2011], tabela 1, [Kemann, Kleppe, Scagliola 2014].

[15] Por. m.in.: [Kemann, Kleppe, Scagliola 2014] - w próbie 288 naukowców częstotliwość użytkowania Google i Google Scholar na skali od 1 do 5 wynosiła odpowiednio (dominanta, mediana): 5, 4,89 i 5, 3,53. [Werla 2013]: w pierwszym kwartale 2013 roku Google wraz z pozostałymi wyszukiwarkami generowało prawie 40% przekierowań na stronę Wielkopolskiej Biblioteki Cyfrowej. [De Rosa et al. 2005]: 84% respondentów rozpoczyna wyszukiwanie za pośrednictwem Google lub innej wyszukiwarki, zaś tylko 1% - na stronie biblioteki uniwersyteckiej. Wśród respondentów będących studentami wartości te wynoszą odpowiednio 89% i 2%. Obszerniejsza bibliografia tego tematu: patrz np. [Orduna-Malea, López-Cózar 2005]

[16] Por. problemy repozytoriów z widocznością opisywane np. w [Orduna-Malea, López-Cózar 2005], [Arlitsch, O'Brien 2012].

[17] Zob. FAQ: *Why does OARR compare institutional and disciplinary open access repositories?* In. *2014 Open Access Repository Ranking* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://repositoryranking.org/?page_id=997.

[18] Por. [Arlitsch, O'Brien 2013], s. 23-30.

[19] [Arlitsch, O'Brien 2013], ramka na s. 27, także s. 53-54.

[20] Zob. np. [Page et al. 1998], s. 3-5, s. 10-11.

[21] Jakakolwiek próba cytowania standardów W3C w dziedzinie architektury stron internetowych byłaby ogromnym zadaniem - zob. *Metadata Task Force Report. W3C Editor's Draft 20 September 2014*. In *World Wide Web Consortium* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://w3c.github.io/dpub-metadata/.

[22] [Arlitsch, O'Brien 2013], s. 85.

[23] [Arlitsch, O'Brien 2013], s. 73.

[24] [Arlitsch, O'Brien 2013], s. 27.

[25] [Arlitsch, O'Brien 2012], s. 61, s. 67, s. 73-76.

[26] Doskonały przykład współpracy między Twitterem a repozytorium akademickim podała Jane Tinkler w swoim wystąpieniu *Open Access + Social Media = Increased Downloads* zaprezentowanym w ramach konferencji REF 2014. W załączonej tabeli (Rys. 3) widać doskonale korelację między kolejnymi „ćwierknięciami na Twitterze oraz wzmiankami w blogosferze a zwiększoną ilością odwiedzin konkretnego artykułu w repozytorium.

[27] [Werla 2013], s. 8.

Bibliografia:

[1] Adamick J., Reznik-Zellen R., *Representation and recognition of subject repositories*, "D-Lib Magazine", [online], 2010, Vol. 16, nr 9/10 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.dlib.org/dlib/september10/adamick/09adamick.html.

[2] Aguillo I. F. [et al.], *Indicators for a webometric ranking of Open Access repositories*, "Scientometrics", [online], 2010, Vol. 82, nr 3, s. 477-486 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://digital.csic.es/bitstream/10261/32190/1/Ranking%20of%20Repositories.pdf.

[3] Arlitsch K., O'Brien P. S., *Improving the visibility and use of digital repositories through SEO* [online], 2013 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://books.google.pl/books?hl=pl&lr=&id=KxKSAwAAQBAJ.

[4] Arlitsch K., O'Brien P. S., *Invisible institutional repositories : addressing the low indexing ratios of IRs in Google Scholar,* "Library Hi Tech", [online], 2012, Vol. 30, nr 1, s. 60-81 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://scholarworks.montana.edu/xmlui/bitstream/handle/1/3193/Arlitsch-Obrien-LHT-GS-final-revised_2012-02-18.pdf.

[5] Bergman M. K., *White Paper: the deep Web: surfacing hidden value,* "Journal of Electronic Publishing", [online], 2001, Vol. 7, nr 1 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://quod.lib.umich.edu/cgi/t/text/idx/j/jep/3336451.0007.104/--white-paper-the-deep-web-surfacing-hidden-value?rgn=main;view=fulltext.

[6] Björk B. C. [et al.], *Anatomy of green Open Access*, "Journal of the American Society for Information Science and Technology", [online], 2014, Vol. 65, nr 2, s. 237-250 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.openaccesspublishing.org/apc8/Personal%20VersionGreenOa.pdf.

[7] Björk B. C. [et al.], *Open Access to the scientific journal literature: situation 2009*. In *PLOS ONE* [online], 2010 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0011273.

[8] Calderón Martinez A., Ruiz Conde E., *The participation and Web visibility of university digital repositories in the European context*, "Comunicar: Media Education Research Journal", 2013, Vol. 20, nr 40, s. 193-200.

[9] *COAR - Annual Report 2013/14*. In *Confederation of Open Access Repositories* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: https://www.coar-repositories.org/files/COAR-Annual-Report-2013-14_public.pdf.

[10] Cothran T., *Google Scholar acceptance and use among graduate students : a quantitative study*, "Library and Information Science Research", [online], 2001, Vol. 33, nr 4, s. 293-301 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.sciencedirect.com/science/article/pii/S0740818811000594.

[11] De Rosa C. [et al.], *Perceptions of libraries and information resources : a report to the OCLC Membership* [online], 2005 [dostęp: 2014-11-02]. Dostępny w World Wide Web: https://oclc.org/content/dam/oclc/reports/pdfs/Percept_all.pdf.

[12] Eve M., Schwarz B., *Q&A : Martin Eve on why we need a public library of the humanities and social sciences*, "Library Journal", [online], 2013, January 15 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://lj.libraryjournal.com/2013/01/oa/qa-martin-eve-on-why-we-need-a-public-library-of-the-humanities-and-social-sciences/.

[13] Gentil-Beccot A., Mele S., Brooks T. C., *Citing and reading behaviours in high-energy physics : how a community stopped worrying about journals and learned to love repositories*, "Scientometrics", [online], 2009, Vol. 84, nr 2, s. 345-355 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://arxiv.org/ftp/arxiv/papers/0906/0906.5418.pdf.

[14] Hatala M. [et al.], *The interoperability of learning object repositories services : standards, implementations and lessons learned*. In *Proceedings of the 13th International World Wide Web conference on Alternate track papers & posters* [online], 2004, s. 19-27 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://dl.acm.org/citation.cfm?id=1013371.

[15] *Inclusion guidelines for webmasters*. In *Google Scholar* [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.google.com/intl/en/scholar/inclusion.html.

[16] Jacsó P., *Google Scholar : the pros and the cons,* "Online Information Review", [online], 2006, Vol. 29, nr 2, s. 208-214 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.researchgate.net/publication/220207633_Google_Scholar_the_pros_and_the_cons/file/3deec528451d9a5d5a.pdf.

[17] Jacsó P., *Google Scholar revisited*, "Online Information Review", [online], 2008, Vol. 32, nr 1, s. 102-114 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://cs.unibo.it/~cianca/wwwpages/dd/08Jacso.pdf.

[18] Kemann M., Kleppe M., Scagliola S., *Just Google it – digital research practices of humanities scholars*. In Mills C., Pidd M., Ward E., *Proceedings of the Digital Humanities Congress 2012 : Studies in the Digital Humanities*, [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://arxiv.org/abs/1309.2434.

[19] Laakso M., Björk B. C., *Anatomy of open access publishing : a study of longitudinal development and internal structure*, "BMC Medicine", [online], 2012, Vol. 10, October, art. nr 124 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.biomedcentral.com/1741-7015/10/124.

[20] *Metadata Task Force Report. W3C Editor's Draft 20 September 2014*. In *World Wide Web Consortium* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://w3c.github.io/dpub-metadata/.

[21] *Open Access Repository Ranking* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://repositoryranking.org.

[22] Orduna-Malea E., López-Cózar E. D., *The dark side of Open Access in Google and Google Scholar : the case of Latin-American repositories* [online], 2005 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://arxiv.org/abs/1406.4331.

[23] Ortega J. L., Aguillo I. F., *Microsoft Academic Search and Google Scholar Citations : a comparative analysis of author profiles*, "Journal of the Association for Information Science and Technology", [online], 2014, Vol. 65, nr 6, s. 1149-1156 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://jlortega.scienceontheweb.net/articles/Ortega_Aguillo_MAS_GSC.pdf.

[24] Page L. [et al.], *The PageRank citation ranking : bringing order to the Web*. In *Stanford InfoLab* [online], 1998 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

[25] Pinfield S. [et al.], *Open-access repositories worldwide, 2005-2012 : past growth, current characteristics and future possibilities*, "Journal of the Association for Information Science and Technology", [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://onlinelibrary.wiley.com/doi/10.1002/asi.23131/full.

[26] Szprot J. (red.), *Otwarta nauka w Polsce 2014 : diagnoza* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://pon.edu.pl/index.php/nasze-publikacje?pubid=13.

[27] Thomas C., McDonald R. H., *Measuring and comparing participation patterns in digital repositories*, "D-Lib Magazine", [online], 2007, Vol. 13, nr 9/10 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.dlib.org/dlib/september07/mcdonald/09mcdonald.html.

[28] Tinkler J., *Open Access + social media = increased downloads*. In *YouTube* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: https://www.youtube.com/watch?v=inYzQABuJ-Y.

[29] *Użyteczność serwisu i analiza SEO – droga do sukcesu*. In *Centrum Projektów Informatycznych* [online], 2013 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://www.cpi.gov.pl/article,uzytecznosc_serwisu_i_analiza_seo___droga_do_sukcesu_,354.html.

[30] Werla M., *Dobre praktyki udostępniania on-line baz bibliograficznych i pełnotekstowych*. In *Materiały Konferencyjne EBIB* [online], 2013, Nr 24 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://open.ebib.pl/ojs/index.php/Mat_konf/article/viewFile/43/166.

[31] Van de Velde E., *The metadata bubble*. In *SciTechSociety* [online], 2014 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://scitechsociety.blogspot.com/2014/10/the-metadata-bubble.html.

[32] Xia J., *A comparison of subject and institutional repositories in self-archiving practices*, „The Journal of Academic Librarianship , [online], 2008, Vol. 34, nr 6, s. 489-495 [dostęp: 2014-11-02]. Dostępny w World Wide Web: http://arizona.openrepository.com/arizona/bitstream/10150/105552/1/Self-archiving.pdf.

Informacja o autorach:

Tomasz Lewandowski – członek zespołu Platformy Otwartej Nauki (http://pon.edu.pl/); zajmuje się kontaktami z wydawcami i redakcjami czasopism udostępnianych w Bibliotece Nauki; tel. (22) 874 91 61, e-mail: t.lewandowski@icm.edu.pl.

Michał Starczewski - członek zespołu Platformy Otwartej Nauki (http://pon.edu.pl/); zajmuje się analizą systemu komunikacji naukowej oraz kontaktami z wydawcami zainteresowanymi udostępnieniem czasopism w Bibliotece Nauki; tel. (22) 874 94 66, e-mail: m.starczewski@icm.edu.pl.