Homo Ludens 1/(3) (2011)

Podsumowanie roku Homo Ludens *1/2009 w internecie*

JAKUB MARSZAŁKOWSKI
Politechnika Poznańska

Abstract

Summary of a year of *Homo Ludens* 1/2009 online

By decision of the editors, the full text of Homo Ludens was made available on the Internet free of charge as PDF files downloadable from the pages of the periodical. Since it seems that most scientific journals still prefer the printed medium, and those that have electronic versions are usually available in closed and paid form, special attention should be paid to the effects of this decision. The year that elapsed is a good period for collecting data on the downloads – data from which one can now gather statistics and draw conclusions. The paper explains the methodology of data collection and presents aggregated statistics of these. Particularly noteworthy are the downloads from abroad, including the ones from carefully examined international scientific bodies. The findings and conclusions presented can definitely be applied to other online journals that are in doubt over the decision to become open and unpaid.

28 października 2009 roku pierwszy numer *Homo Ludens*, wydawany przede wszystkim w tradycyjnej wersji drukowanej, decyzją redakcji został również udostępniony pełnotekstowo i nieodpłatnie w Internecie – zarówno w posta-

ci pojedynczych artykułów, jak i całego czasopisma (w jednym pliku). Jako że standardem publikacji elektronicznej artykułów i wydawnictw naukowych jest format PDF[1], został on wybrany również tutaj. W technologii PHP+MySQL[2] stworzona została dość prosta strona internetowa pozwalająca na pobieranie tekstów, wzbogacona jedynie o licznik pobrań dla każdego pliku. Strona funkcjonuje pod adresem http://ptbg.org.pl/HomoLudens/. Licznik pobrań przez rok obecności pierwszego numeru *Homo Ludens* w Internecie zebrał potężną ilość danych, pozwalającą na stworzenie niniejszego podsumowania.

1. Zbierane informacje

Stworzona na potrzeby licznika baza danych odnotowywała adresy IP[3] pobierających[4], wraz ze wskaźnikiem czasu i indeksem pobieranego pliku. W wersji pierwotnej miało to służyć wyłącznie eliminacji naliczania tego samego pobrania więcej niż jeden raz. Dodatkowo strona *Homo Ludens*, jak wszystkie strony PTBG, została podłączona do analizatora ruchu sieciowego Google Analytics (Google, 2010).

Z perspektywy roku, ogromnej liczby pobrań i pracy włożonej w przygotowanie tego raportu oczywistym wnioskiem jest, że od początku należało zbierać więcej informacji dostępnych w nagłówkach HTTP[5] (Network Working Group, 1999), w polach takich jak:

User-Agent – pole to zawiera ciąg znakowy, którym przedstawia się klient[6] (może to identyfikować przeglądarkę internetową albo crawlera);

Accept-Language – zawiera listę języków akceptowanych przez klienta; zaczyna się ona od języka przeglądarki internetowej, który zazwyczaj jest językiem ojczystym internauty;

Referer – zawiera adres strony HTML, z której nastąpiło przekierowanie do obecnej strony lub pliku.

[1] PDF (ang. *Portable Document Format*) – otwarty standard zapisu dokumentów zapewniający ich przenośność. W założeniach dokument można otworzyć na każdej platformie i wszędzie będzie on wyglądał tak samo.

[2] PHP to skryptowy język programowania, MySQL to system zarządzania relacyjnymi bazami danych. Razem stanowią najpopularniejszą platformę dla tworzenia aplikacji internetowych.

[3] Adres IP to liczbowy adres nadawany elementom, w tym komputerom, tworzącym sieci – przede wszystkim Internet. Składa się z czterech liczb z zakresu 0-255 (8 bitów) oddzielonych kropkami, np. 127.0.0.1.

[4] Wszystkie dane gromadzone są wyłącznie dla celów naukowych i statystycznych.

[5] HTTP (ang. *Hypertext Transfer Protocol*) to protokół tworzący sieć WWW, zapewniający transfer stron HTML, obrazków, plików.

[6] Klient to program lub system, który podłącza się przez sieć do komputera lub systemu będącego serwerem.

Zostało to uzupełnione w najnowszej wersji strony wraz z drugim numerem *Homo Ludens*. Pewne ogólne wnioski wyciągane na podstawie niewielkiej próby danych zebranej od tego czasu będą przywoływane dalej w tym raporcie.

2. Crawlery

Zawsze wśród pobrań plików lub stron internetowych część stanowią te dokonywane przez crawlery. Crawler, spider, bot, po polsku czasem pełzacz, to komputer automatycznie przeglądający i indeksujący zasoby Internetu, zazwyczaj na potrzeby jakiejś wyszukiwarki. Najprostszym sposobem rozróżnienia takiego bota jest sprawdzenie ciągu identyfikacyjnego klienta User-Agent – wszystkie popularne przedstawiają tam swoje dane, zaś kilka próbujących się ukrywać generuje na tyle niewielki ruch, że można je bez wahania pominąć. Ponieważ w badanym okresie strony *Homo Ludens* nie zbierały informacji z pola User-Agent, konieczna była identyfikacja crawlerów na podstawie adresów IP, nieco bardziej pracochłonna, ale również możliwa. Znaleziono następujące pobrania:

googlebot, crawler indeksujący dla Google – 211 adresów IP, łącznie 1539 pobrań;

msnbot, crawler microsoftu indeksujący m.in. dla Bing – 292 adresów IP, łącznie 858 pobrań;

crawl yahoo, crawler wyszukiwarki Yahoo – 28 adresów IP, łącznie 552 pobrań;

baiduspider, crawler chińskiej wyszukiwarki Baidu – 12 adresów IP, łącznie 14 pobrań;

yandex, cuil, ask, inne crawlery – łącznie 15 adresów IP i 151 pobrań.

Powyższe zestawienie wskazuje na to, że artykuły zostały zaindeksowane przez wszystkie przeglądarki i pojawiają się w ich wynikach wyszukiwania.

Analiza fragmentarycznych danych ze wzbogaconej wersji strony nie wykazała, by jakieś crawlery zostały pominięte w wyniku analizowania adresów IP; jednocześnie sugeruje ona faktyczne pobrania plików za pomocą wyżej wymienionych wyszukiwarek.

3. Ogólne statystyki

Po usunięciu (zgodnie z metodologią opisaną w poprzedniej sekcji) tych danych o pobraniach, które dotyczyły crawlerów, przeprowadzono – metodą geolokacji IP – analizę informacji o kraju pochodzenia osób pobierających

artykuły. W jednym z prostszych wydań polega ona na wykorzystaniu tablicy z zakresami adresów IP i ich przypisaniami geograficznymi (Siwpersad, Gueye, & Uhlig, 2008). Przyjmuje się, że metody te, w zależności od jakości zastosowanej tablicy, pozwalają nawet dość dokładnie przypisywać adresy IP do regionów czy miast, a w skali całych krajów są wystarczające ponad wszelką wątpliwość.

Uzyskane informacje kształtują się w sposób następujący:

- 14 005 pobrań plików z Polski (z 11 636 unikatowych adresów IP);
- 1252 pobrania plików z zagranicy (z 827 różnych adresów IP z 58 różnych krajów, w kolejności liczby pobrań: Stany Zjednoczone, Niemcy, Wielka Brytania, Rosja, Ukraina, Chiny, Francja, Kanada, Irlandia, Japonia, Norwegia, Szwecja, Austria, Senegal, Włochy, Hiszpania, Czechy, Holandia, Izrael, Jugosławia, Dania, Rumunia, Indonezja, Singapur, Belgia, Turcja, Portugalia, Kolumbia, Białoruś, Słowacja, Litwa, Grecja, Australia, Finlandia, Brazylia, Kazachstan, Filipiny, Szwajcaria, Łotwa, Malezja, Indie, Iran, Republika Korei, Chile, Tajlandia, Tajwan, Tunezja, Luksemburg, Bośnia i Hercegowina, Meksyk, Cypr, Bułgaria, Islandia, Republika Południowej Afryki, Bahamy, Azerbejdżan, Egipt, Oman).

W tym samym okresie strony PTBG (cały serwis ptbg.org.pl) odwiedziło 14 487 unikatowych użytkowników[7] – w tym stronę główną *Homo Ludens*, zawierającą wszystkie linki do plików .pdf do pobrania, zaledwie 2996[8]. Wobec przeszło dwunastu tysięcy adresów IP, które odnotowano w pobraniach artykułów, możliwe jest postawienie tylko jednego wniosku. Skoro dwanaście tysięcy osób nie dotarło do tekstów za pośrednictwem strony głównej, większość pobrań została dokonana bezpośrednio z wyników wyszukiwania w wyszukiwarkach internetowych. Te ostatnie wyświetlają fragmenty tekstu z plików .pdf (zawierające szukane frazy) oraz linkują bezpośrednio do tych plików, z pominięciem stron, na których zostały one zamieszczone. Analiza fragmentarycznych danych ze wzbogaconej wersji strony potwierdza te wnioski, jednocześnie wskazując na ogromną dominację Google (mieści się w tym pewien – trudny do określenia – udział Google Scholar, wyszukiwarki prac naukowych) i tylko pojedyncze wystąpienia innych wyszukiwarek. Odpowiada to sytuacji ogólnej, w której Google obsługuje niezmiennie ponad 97% wyszukiwań w Polsce (Gemius, 2010), a na świecie mniejsze udziały ma tylko na kilku lokalnych rynkach (Rosja, Chiny, Czechy).

[7] Unikatowy użytkownik, w skrócie UU, lub czasem: unikatowy odwiedzający – to możliwy do oznaczenia różnymi metodami z różną precyzją niepowtarzalny użytkownik Internetu. Więcej na ten temat: Marszałkowski, 2010.

[8] Dane z Google Analytics (Google, 2010).

4. Pobrania z zagranicy

Dla wszystkich zagranicznych adresów IP wykonana została odwrotna translacja adresów DNS[9], zamieniająca adres IP z formy cyfrowej, zrozumiałej dla maszyn, na formę domenową, tekstową, a więc zrozumiałą dla ludzi (LeFebvre, Craig, 1999). Metoda ta nie zapewnia stuprocentowej skuteczności – ponad dwustu adresów nie udało się odwrócić.

Otrzymane adresy DNS zostały poddane analizie. Przeważały wśród nich adresy charakterystyczne dla funkcjonujących w danych krajach dostawców Internetu wykorzystujących sieci telefoniczne, kablowe, komórkowe – adresy możliwe do rozpoznania po nazwach dostawców (orange, tele2, vodafone, bt etc.) lub nazwach usług i technologii (adsl, cable, dyn, broadband, ppp...). Ponadto stwierdzono adresy firmowe kilkunastu wiodących światowych koncernów.

Ponieważ wstępny przegląd listy domen pozwolił dostrzec, że pobrań *Homo Ludens* dokonywano również z sieci akademickich, podjęta została decyzja o dokładniejszym przeanalizowaniu tej części pobrań. Identyfikacja większości z nich była możliwa dzięki charakterystycznym adresom domenowym, zgodnym z przyjętą w kilku krajach systematyką. I tak domena .edu oznacza amerykańskie jednostki naukowe (jak berkeley.edu), a domena .ac.uk – brytyjskie ośrodki akademickie (jak ox.ac.uk). Subdomena ac. jest też używana w kilku innych krajach, choć nie w tak systemowy sposób – znaleziono jednak w ten sposób uczelnie belgijskie, japońskie, ukraińskie. Uniwersytety niemieckie używają bardzo często w nazwie przedrostka uni- (jak uni-leipzig.de), choć także nie jest to powszechne. Wreszcie pozostałe uczelnie zostały oznaczone metodą ręcznego sprawdzenia wyodrębnionej podgrupy adresów, które mogły być adresami akademickimi.

Ostateczna lista uczelni, z których pobierano artykuły z *Homo Ludens*, wraz z podziałem na kraje ma postać:

- **Niemcy** (12) Uniwersytet w Lipsku, Uniwersytet Justusa Liebiga w Giessen, Uniwersytet Eberharda Karola w Tybindze, Uniwersytet Jana Gutenberga w Moguncji, Uniwersytet im. Georga Augusta w Getyndze, Uniwersytet Ruhry w Bochum, Europejski Uniwersytet Viadrina we Frankfurcie nad Odrą, Uniwersytet Ottona von Guerickego w Magdeburgu, Uniwersytet Techniczny w Regensburgu, Uniwersytet Technicz-

[9] DNS (ang. *Domain Name System*) – system nazw serwerów w postaci takiej jak widoczna w polu przeglądarki.

ny w Chemnitz, Uniwersytecie Ludwiga Maksymiliana w Monachium, Europejska Akademia Środowiska Miejskiego w Berlinie.

- **Wielka Brytania** (6) Uniwersytet Oxfordzki, St Edmund's College w Cambridge, Imperial College w Londynie, Anglia Ruskin University w Cambridge, Uniwersytet w Liverpoolu, University of Central Lancashire w Preston, Woodham Community Technology College
- **Stany Zjednoczone** (3) Uniwersytet Kalifornijski w Berkeley, Uniwersytet Kalifornijski w Los Angeles, Uniwersytet Wirginii w Charlottesville
- **Czechy** (3) Uniwersytet Karola w Pradze, Uniwersytet Masaryka w Brnie, Wyższa Szkoła Ekonomiczna w Pradze
- **Japonia** (2) Uniwersytet Osakijski, Uniwersytet Tsukuba
- **Kanada** (2) Uniwersytet Quebec w Trois Rivieres, Kanadyjski Państwowy Komitet Badań Naukowych
- **Ukraina** (2) Narodowa Akademia Nauk Ukrainy w Doniecku, Lwowska Biblioteka Naukowa
- **Belgia** Wolny Uniwersytet Brukselski
- **Austria** Uniwersytet Nauk Stosowanych Joanneum w Grazu
- **Norwegia** Uniwersytet w Oslo
- **Szwajcaria** Europejska Organizacja Badań Jądrowych CERN
- **Portugalia** Uniwersytet w Minho
- **Grecja** Uniwersytet Demokryta w Tracji
- **Chile** Katolicki Uniwersytet Chile w Santiago

Należy tu jednocześnie podkreślić, że teoretycznie taki adres publiczny, jaki jest odnotowywany przy pobraniu, może wskazywać zarówno konkretny komputer, jak i podsieć (pracownię, budynek itd.) z pulą adresów prywatnych wewnątrz, ale z zewnątrz widoczną jako pojedynczy adres[10]. W praktyce jednak odwrócone adresy DNS dla uczelni były na tyle dokładne, że można było określać dokładniej, skąd w strukturze danej uczelni dokonano pobrania. Czasem adresy wskazywały na komputery do użytku studentów (np. w *white space*[11]), czasem na konkretne instytuty (np. slawistyka na jednym z niemieckich uniwersytetów), a czasem nawet można byłoby określić konkretne biurko i komputer, przypisane do określonego naukowca.

[10] Na przykład z wykorzystaniem technologii NAT (ang. *Network Address Translation*). W zasadzie nie ma ograniczenia liczby komputerów w takiej podsieci, w praktyce spotyka się nawet sieci osiedlowe złożone z kilku bloków ukryte za pojedynczym adresem IP.

[11] Wyznaczone na uczelni pomieszczenia, często otwarta przestrzeń, ze stanowiskami komputerowymi przeznaczonymi do swobodnego wykorzystania przez studentów.

5. Pobrania z Polski

Pula adresów IP z Polski była zbyt duża dla przeprowadzenia tak dokładnej analizy, jak w przypadku adresów zagranicznych. Już odwrotna translacja adresów DNS ponad jedenastu tysięcy adresów zajęłaby bardzo dużo czasu – proces ten nie należy do najszybszych, zwłaszcza dla adresów, które trudno jednoznacznie zakwalifikować. Dalej analiza otrzymanej puli adresów domenowych przy tej objętości również byłaby bardzo trudna. Dlatego do sprawdzenia wybrana została lista 128 adresów IP, z których dokonano największej liczby pobrań (powyżej pięciu plików). Celem było wyszukanie ewentualnych wad zgromadzonych danych.

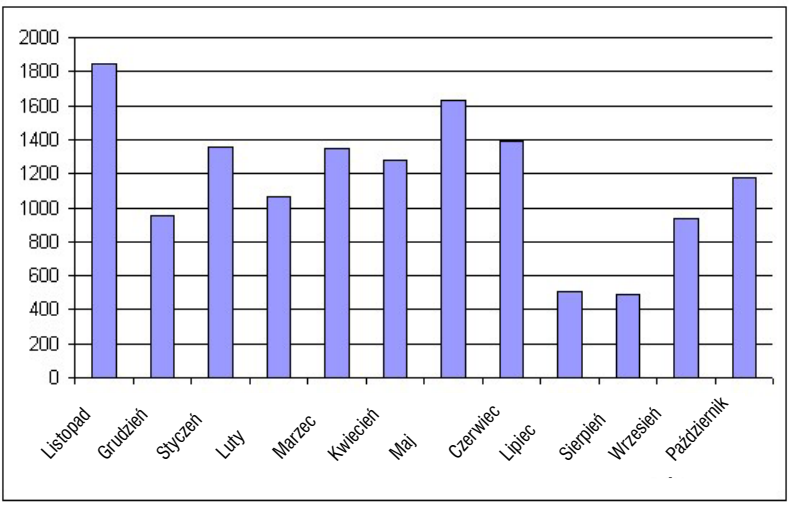
Można przyjąć, że w Polsce właściwie nie ma crawlerów. Jedyny powszechnie znany, NetSprint zbierający dane dla wyszukiwarki Wirtualnej Polski[12], znajdował się na wspomnianej liście z jednym adresem IP i dwudziestoma dwoma pobraniami plików. Ponadto stwierdzono adresy kilkunastu krajowych uczelni i adresy charakterystyczne dla wszystkich wiodących dostawców Internetu. Założono, że z minimalnym błędem można przyjąć dane dla Polski jako złożone w pełni z poprawnych pobrań.

6. Wnioskowanie z zebranych danych

Dane zbierane w powyższej metodologii nie dają odpowiedzi na kilka pytań: Czy artykuł został po pobraniu przeczytany? Czy został wykorzystany w pracy naukowej? Czy język polski nie okazał się problemem w przypadku pobrań z zagranicy?

Wydaje się, że jedynym sposobem na udzielenie całościowej odpowiedzi na dwa pierwsze pytania jest śledzenie cytowań. To jednak wciąż otwarty, trudny problem naukowym z pogranicza bibliografii i informatyki. Istniejące rozwiązania są realizowane przez duże podmioty, dużym kosztem i cały czas nie rozwiązują całości problemu. Należy więc uznać, że jest to w tej chwili poza możliwościami Polskiego Towarzystwa Badania Gier. Jednocześnie okres jednego roku – przy obecnych cyklach wydawniczych czasopism naukowych – to zdecydowanie za mało, by możliwe było wskazanie pojedynczych choć cytowań.

[12] NetSprint obsługuje poniżej 2% wyszukań w Polsce (Gemius, 2010).



Wykres 1. Pobrania w czasie (miesiące)

Można jednak próbować wnioskowania pośredniego z innych dostępnych danych. Wykres 1 przedstawia rozkład pobrań w czasie. W listopadzie, świeżo po opublikowaniu pierwszego numeru *Homo Ludens* w Internecie, liczba pobrań była najwyższa, ale są to zapewne w ogromnej większości sami członkowie PTBG. Poza listopadem szczyt liczby pobrań przypada na maj i czerwiec, szczyt studenckiej gorączki związanej z pisaniem prac dyplomowych. Następnie dwa miesiące, wakacje w świecie akademickim, to niemalże trzykrotny spadek liczby pobrań, od września znów ich przybywa – być może to studenci kończący pisanie prac dyplomowych w drugim terminie. Dane te mogą wskazywać w znacznej części akademickie wykorzystanie pobranych artykułów.

Oczywiste jest, że abstrakty i tytuły w języku angielskim mogły powodować pojawianie się artykułów w wyszukiwaniach angielskojęzycznych i przyczynić się do części pobrań z zagranicy; później jednak polski język publikacji mógł okazać się barierą nie do przejścia. Tu jednak także można próbować wyciągnąć pewne wnioski nie wprost. Wśród zagranicznych uczelni przeważają niemieckie, w tej liczbie zaś można wskazać typowe niemieckie uniwersytety, gdzie polscy studenci jadą na wymianę w ramach programu Erasmus. Z kolei na przykład artykuł „Gry, procedury, przewidywalność – w kontekście polsko-japońskiej komunikacji międzykulturowej" Arkadiusza Jabłońskiego mógł spowodować zainteresowanie kogoś z uczelni japońskich i jego obecność w periodyku czyni te pobrania łatwiejszymi do wyjaśnienia. Z pewnością możliwe byłoby wyciąganie dalszych takich wniosków.

Należy tu również podkreślić, że w przypadku papierowych wydań czasopism naukowych egzemplarz musi zostać nabyty, zanim możliwe będzie stwierdzenie, czy artykuł w nim zawarty przyda się do czegokolwiek w dalszych badaniach. Ani nakład, ani liczba sprzedanych egzemplarzy nie są więc w żadnym stopniu bardziej miarodajne niż liczba pobrań. Tym bardziej więc każde pobranie artykułu należy traktować jak rozpowszechniony jego egzemplarz.

7. Podsumowanie

Bez cienia wątpliwości można stwierdzić, że żyjemy w epoce informacji, a Internet jest ich najpopularniejszym dostarczycielem. Globalna sieć, publikacje elektroniczne, wyszukiwanie i indeksy prac naukowych zmieniają naukę całkowicie. Spędzanie czasu na poszukiwaniach w bibliotekach funkcjonuje już wyłącznie w opowieściach starszych pracowników naukowych, zawsze zakończone komentarzem skierowanym do studentów czy doktorantów, że ci już tak robić nie muszą.

Przełożenie 200 egzemplarzy papierowych, w dodatku nie w całości sprzedanych, na przeszło 15 000 pobrań, w tym ponad 1200 z zagranicy, musi robić wrażenie. W dodatku są to wyniki z jednego tylko roku, a artykuły będą nadal pobierane w latach kolejnych. To wskazuje, że decyzja redakcji o udostępnieniu wszystkich artykułów w wersji pełnotekstowej nieodpłatnie w Internecie była bardzo trafna.

Wydaje się też, że zaprezentowane wyniki pozwalają sformułować zalecenie dla wszystkich wydawnictw naukowych publikowanych cały czas tylko na papierze, by jak najszybciej wzbogaciły się o wydania elektroniczne publikowane w Internecie. Jeżeli uznałyby za niewłaściwe bezpłatne udostępnianie artykułów pełnotekstowych, to mogą to uczynić w jednym z wielu systemów pozwalających na dostęp wyłącznie za opłatą. Nadal jednak teksty takie będą widoczne z poziomu wszystkich wyszukiwarek oraz natychmiastowo dostępne dla każdego, kto będzie chciał z nich korzystać.

LITERATURA

Gemius. (2010). *gemiusranking: Wyszukiwarki — silniki*. Online: <http://www.ranking.pl/pl/rankings/search-engines.html>.

Google. (2010). *Google analytics — official website*. Online: <http://www.google.com/analytics/>.

LeFebvre, W., Craig, K. (1999). Rapid reverse dns lookups for web servers. In: *Proceedings of the*

2nd conference on usenix symposium on internet technologies and systems - volume 2 (p. 21–21). USENIX Association.

Marszałkowski, J. (2010). Problematyka pomiaru popularności gier przeglądarkowych jako przykładu serwisów internetowych. *Homo Ludens*, *2*, 97-106. Online: <http://ptbg.org.pl/dl/44/>.

Network Working Group. (1999). Hypertext transfer protocol – http/1.1 [Computer software manual]. Online: <http://tools.ietf.org/html/rfc2616>.

Siwpersad, S., Gueye, B., Uhlig, S. (2008). Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts. In: M. Claypool & S. Uhlig (Eds.), *Passive and active network measurement* (Vol. 4979, p. 11-20). Springer: Berlin / Heidelberg.

Data dostępu do źródeł internetowych wykorzystanych w pracy: 30 grudnia 2010.

mgr inż. Jakub Marszałowski, twórca i badacz gier przeglądarkowych, doktorant w Instytucie Informatyki Politechniki Poznańskiej, *jakub.marszalkowski@cs.put.poznan.pl*

Podsumowanie roku *Homo Ludens* 1/2009 w internecie

Abstrakt

Decyzją redakcji czasopismo Homo Ludens zostało pełnotekstowo i nieodpłatnie udostępniono w Internecie w postaci plików PDF pobieralnych ze stron czasopisma. Skoro wydaje się, że większość czasopism naukowych ciągle jeszcze preferuje formę papierową, a te, które mają formy elektroniczne, są dostępne zazwyczaj w formie zamkniętej i płatnej, ze szczególną uwagą należy przyjrzeć się skutkom tej decyzji. Rok, który minął, to dobry okres do zbierania danych na temat pobrań – danych, z których można teraz wyciągnąć wnioski. W artykule wyjaśniona jest metodologia zbierania danych oraz przedstawione są zagregowane z nich statystyki. Na szczególną uwagę zasługują pobrania z zagranicy, w tym dokładnie przeanalizowane pobrania z międzynarodowych instytucji naukowych. Przedstawione wyniki i konkluzje śmiało można odnieść do innych czasopism internetowych, które wahają się z decyzją o otwartości i bezpłatności.