Title: PLOS ONE - studium przypadku analizy cytowań prac naukowych na podstawie danych otwartego indeksu cytowań (OpenCitations Corpus)

Author: Anna Małgorzata Kamińska

Citation style: Kamińska Anna Małgorzata. (2017). PLOS ONE - studium przypadku analizy cytowań prac naukowych na podstawie danych otwartego indeksu cytowań (OpenCitations Corpus). "Biuletyn EBIB" (2017, nr 6).

EBIB Biuletyn EBIB, nr 6 (176)/2017,
Ewaluacja nauki w Polsce
Badania, teorie, opinie

Anna Małgorzata Kamińska
Instytut Bibliotekoznawstwa i Informacji Naukowej
Uniwersytet Śląski w Katowicach
anna.kaminska@us.edu.pl

PLOS ONE – studium przypadku analizy cytowań prac naukowych na podstawie danych otwartego indeksu cytowań (OpenCitations Corpus)

Streszczenie: Artykuł prezentuje studium przypadku obrazujące możliwości prowadzenia analiz bibliometrycznych na podstawie danych otwartego indeksu cytowań nazwanego przez jego twórców OpenCitations Corpus. Dla artykułów cytowanych pochodzących z czasopisma PLOS ONE wyekstrahowano dane z całości korpusu i sformatowano w sposób umożliwiający prowadzenie analiz w narzędziach zewnętrznych (arkusz kalkulacyjny, aplikacja obliczeniowo-wizualizacyjna Gephi). Następnie przeprowadzono przykładowe analizy i wizualizacje grafów cytowań artykułów. Na przykładach zaprezentowano również możliwości języka SPARQL umożliwiającego prowadzenie analiz wprost na platformie OpenCitations udostępnionej jako usługa WWW bądź też uruchomionej we własnym środowisku obliczeniowym.

Słowa kluczowe: OpenCitations, OpenCitations Corpus, indeks cytowań, bibliometria, źródła danych, studium przypadku, Gephi, PLOS ONE

Wprowadzenie

Współczesne trendy dokumentowania badań naukowych na zasadach ich publikowania w czasopismach o otwartym dostępie zaczynają zmieniać stopniowo krajobraz rozwoju dziedzin naukometrycznych w kierunku zwiększenia możliwości prowadzenia badań i rozwoju metod przez badaczy nie tylko związanych bezpośrednio z komercyjnymi dostawcami danych bibliograficznych. Dotychczasowa hegemonia komercyjnych usługodawców w zakresie analiz bibliometrycznych czy udostępniania danych bibliograficznych może zostać zachwiana na skutek przekazywania przez wydawnictwa tych danych nieodpłatnie wszystkim zainteresowanym podmiotom. Lista takich wydawców już jest bardzo pokaźna i wydaje się tylko kwestią czasu, kiedy pozostali ulegną presji wywieranej przez środowiska naukowe. Sytuacja ta stwarza warunki rozwoju niekomercyjnych systemów indeksów cytowań, z których ciekawą propozycją wydaje się OpenCitations Corpus (OCC). Więcej o samej koncepcji, przedsięwzięciu, architekturze składowanych informacji i zastosowanych technologiach oraz ontologiach znaleźć można w opisie samych jej twórców[1] bądź w krajowym artykule przeglądowym autorki[2].

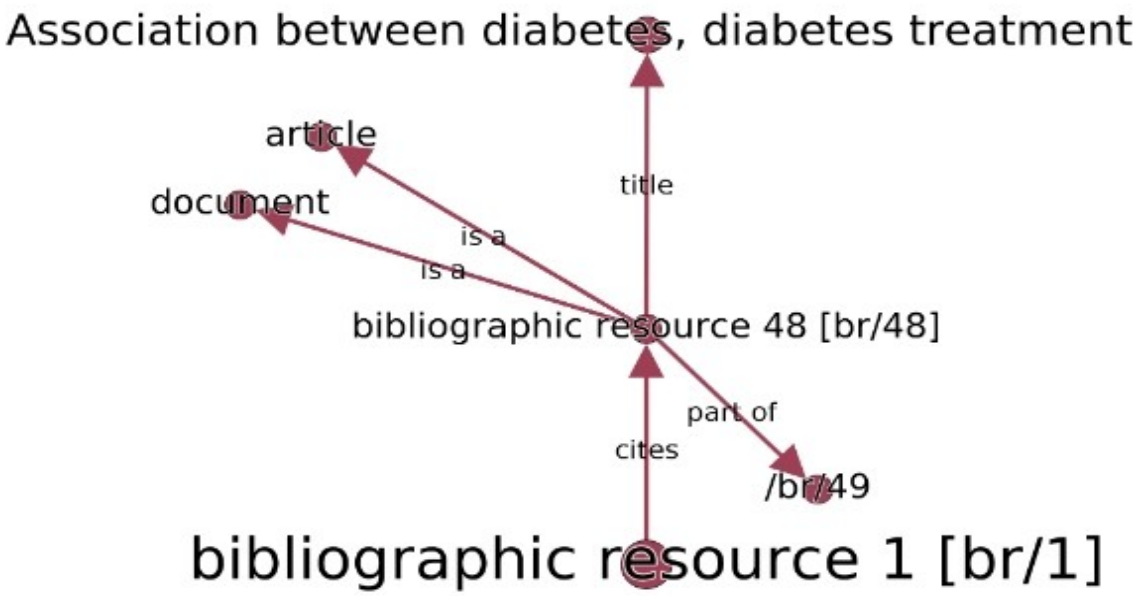
Niniejszy artykuł ma na celu natomiast przedstawienie konkretnego studium przypadku realizacji własnych analiz bibliometrycznych na podstawie danych zaczerpniętych ze wspomnianego korpusu, a dotyczących czasopisma PLOS ONE amerykańskiego wydawcy Pu-

[1] PERONI, S., DUTTON, A., GRAY, T., SHOTTON, D. Setting our bibliographic references free: towards open citation data. *Journal of Documentation* [online]. 2015, 71 (2), s. 253–277. [Dostęp 9.09.2017]. Dostępny w: http://speroni.web.cs.unibo.it/publications/peroni-2015-setting-bibliographic-references.pdf.
[2] KAMIŃSKA, A.M. OpenCitations – otwarty indeks cytowań publikacji naukowych. *Biuletyn EBiB* [online]. 2017, No 176. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/551.

blic Library of Science (PLOS), które razem z brytyjskim BioMed stanowią jedne z największych, jeśli chodzi o publikowanie na zasadach otwartego dostępu.

Korzystanie z zasobów udostępnianych przez OCC znakomicie ułatwia dobrze udokumentowany model pojęciowy (ontologie bibliograficzne), zgodnie z którym zasilana jest baza danych gromadząca informacje w układzie sieciowym (grafowym) w postaci zdań składających się z tzw. trójek reprezentujących kolejno podmiot, orzeczenie i obiekt (przykładowo: <dany artykuł> <jest cytowany> <inny dany artykuł> lub <dany artykuł> <zawiera się> <zeszyt danego czasopisma>). Jest to technika powszechnie używana do definiowania sieci semantycznych, często opisywanych za pomocą formatu RDF[3], zaś wiedza zgromadzona za pomocą takiego opisu może być odkrywana za pomocą języka SPARQL[4], który pozwala na formułowanie przeróżnych zapytań analitycznych.

Celem lepszego zobrazowania zastosowania sieci semantycznych do reprezentacji dziedziny cytowań w oparciu o wybraną ontologię bibliograficzną na rys. 1 przedstawiono wycinek przykładowej sieci.



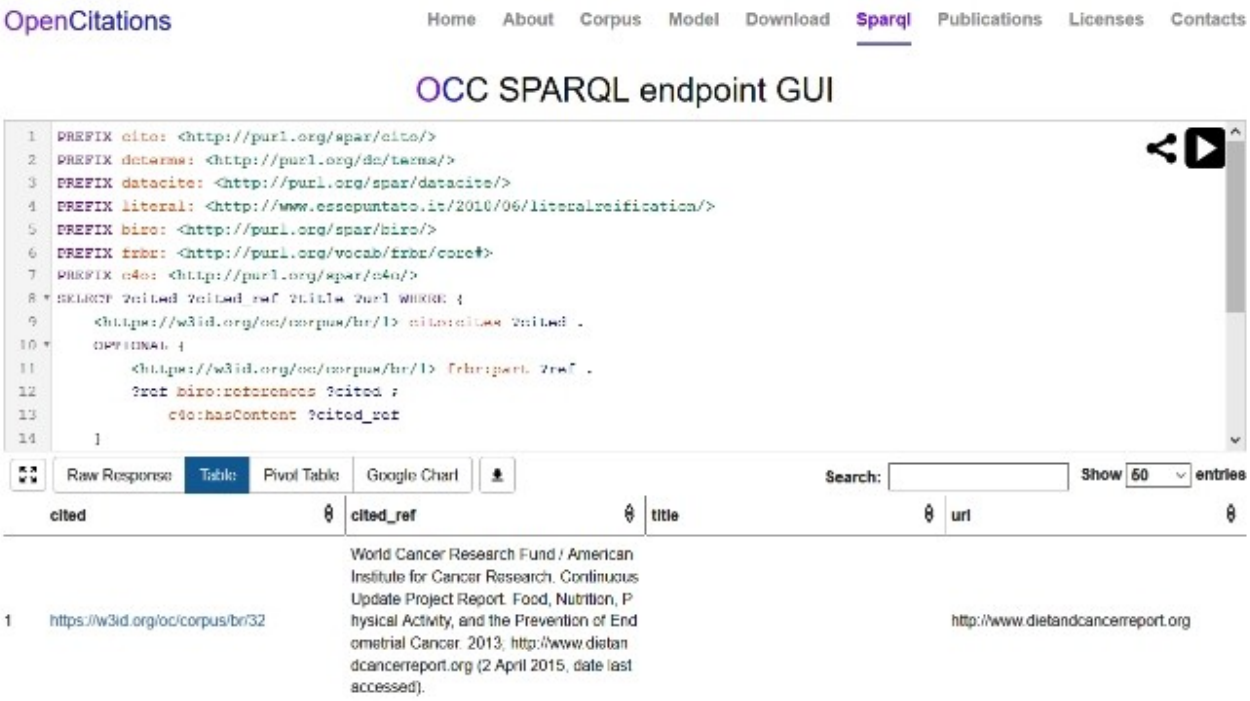
Rys. 1. Wycinek sieci semantycznej opisującej przykładowe informacje bibliograficzne.
Źródło: opracowanie własne.

Widzimy tutaj zasób bibliograficzny [br/1] (ang. *bibliographic resource*), który cytuje (cites) inny zasób bibliograficzny [br/48]. Zasób cytowany opisany jest tytułem (title) oraz typami/klasami (is a), do których należy. Widać tutaj również, że nie jest on samodzielnym bytem, tylko zawarty jest (part of) w innej jednostce bibliograficznej, która jest zeszytem konkretnego czasopisma. Jest to tylko prosty przykład, gdyż typów relacji opisujących (orzeczeń) w tej ontologii jest o wiele więcej.

[3] *RDF* [online]. W3C, 2017. [Dostęp 9.09.2017]. Dostępny w: https://www.w3.org/RDF/.
[4] *SPARQL Query Language for RDF* [online]. W3C, 2017. [Dostęp 9.09.2017]. Dostępny w: https://www.w3.org/TR/rdf-sparql-query/.

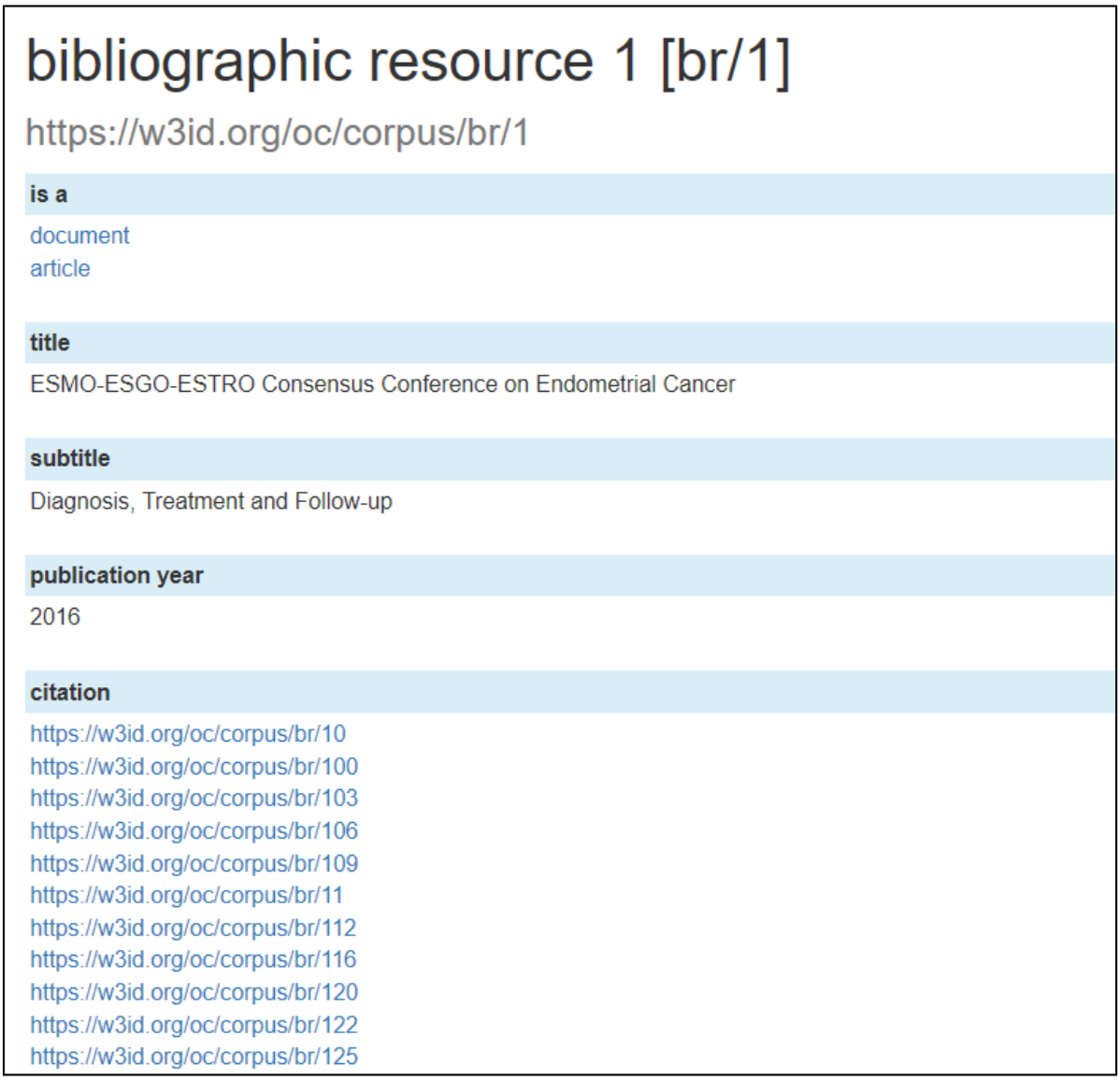
Witryna OpenCitations

Podstawowy wariant analizowania danych polega na korzystaniu wprost z zasobów udostępnianych z poziomu witryny internetowej. Korzystając z zakładki „Sparql" (rys. 2) możliwe jest wysłanie zapytania do systemu źródłowego, a uzyskane odpowiedzi w postaci listy atrybutów pobrać można w jednym z proponowanych formatów wymiany danych.



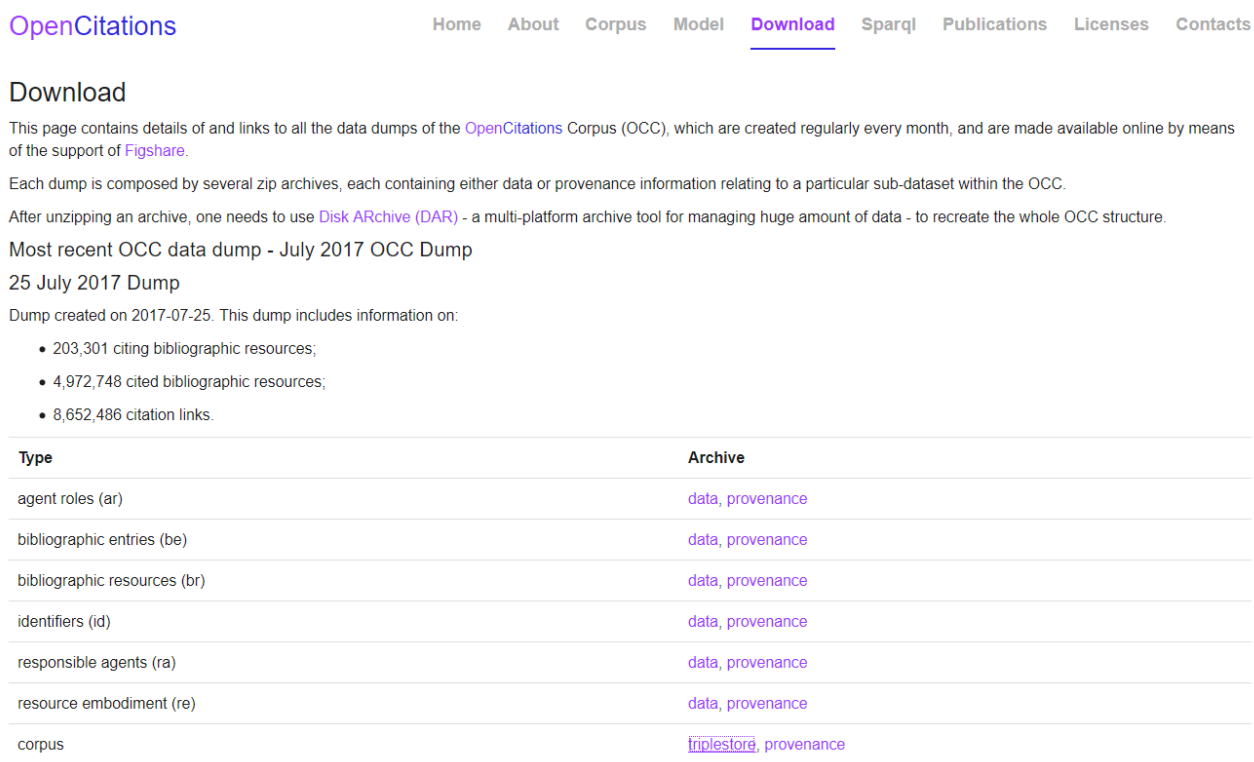
Rys. 2. Zakładka „Sparql" platformy OpenCitations
Źródło: OCC SPARQL endpoint GUI. W: *OpenCitations* [online]. [Dostęp 09.09.2017]. Dostępny w: http://opencitations.net/sparql.

Trzeba jednak zwrócić uwagę, że w chwili obecnej zasoby sprzętowe, na których uruchomione są usługi platformy, są dość skromne, co może powodować wydłużony czas odpowiedzi na zadane pytanie lub nawet całkowite wstrzymanie wykonywanego właśnie zapytania. Dodatkowo pobieranie wyników odpowiedzi zawierających wiele tysięcy rekordów może być mocno kłopotliwe np. z powodu nieoczekiwanego przerwania procesu transmisji pliku zwrotnego. Nie zmienia to jednak faktu, że dla prostych zapytań czy podglądu danych szczegółowych platforma OpenCitations jest w zupełności wystarczająca, a identyfikowanie poszczególnych zasobów zgodne z koncepcją URI (ang. Uniform Resource Identifier) powoduje, że nawigacja z poziomu przeglądarki internetowej po ścieżkach cytowań czy podążanie za jakimikolwiek innymi relacjami są łatwe i intuicyjne (rys. 3).



Rys. 3. Okno podglądu danych szczegółowych dla wybranej jednostki bibliograficznej
Źródło: Bibliographic resource 1. W: *OpenCitations* [online]. [Dostęp 09.09.2017].
Dostępny w: http://opencitations.net/corpus/br/1.html.

Chcąc pobrać całą zawartość korpusu, aktualizowaną w cyklach miesięcznych, należy skorzystać z zakładki „Download", gdzie odnośnik „triplestore" (rys. 4) spowoduje przeniesienie do odpowiedniej strony repozytorium „Figshare", z której pobrać można archiwum ZIP (o aktualnej objętości ponad 20GB), zawierające zarówno dane, jak i oprogramowanie potrzebne do uruchomienia własnej instancji serwera bazy danych.



OpenCitations

Home About Corpus Model Download Sparql Publications Licenses Contacts

Download

This page contains details of and links to all the data dumps of the OpenCitations Corpus (OCC), which are created regularly every month, and are made available online by means of the support of Figshare.

Each dump is composed by several zip archives, each containing either data or provenance information relating to a particular sub-dataset within the OCC.

After unzipping an archive, one needs to use Disk ARchive (DAR) - a multi-platform archive tool for managing huge amount of data - to recreate the whole OCC structure.

Most recent OCC data dump - July 2017 OCC Dump

25 July 2017 Dump

Dump created on 2017-07-25. This dump includes information on:

- 203,301 citing bibliographic resources;
- 4,972,748 cited bibliographic resources;
- 8,652,486 citation links.

Type	Archive
agent roles (ar)	data, provenance
bibliographic entries (be)	data, provenance
bibliographic resources (br)	data, provenance
identifiers (id)	data, provenance
responsible agents (ra)	data, provenance
resource embodiment (re)	data, provenance
corpus	triplestore, provenance

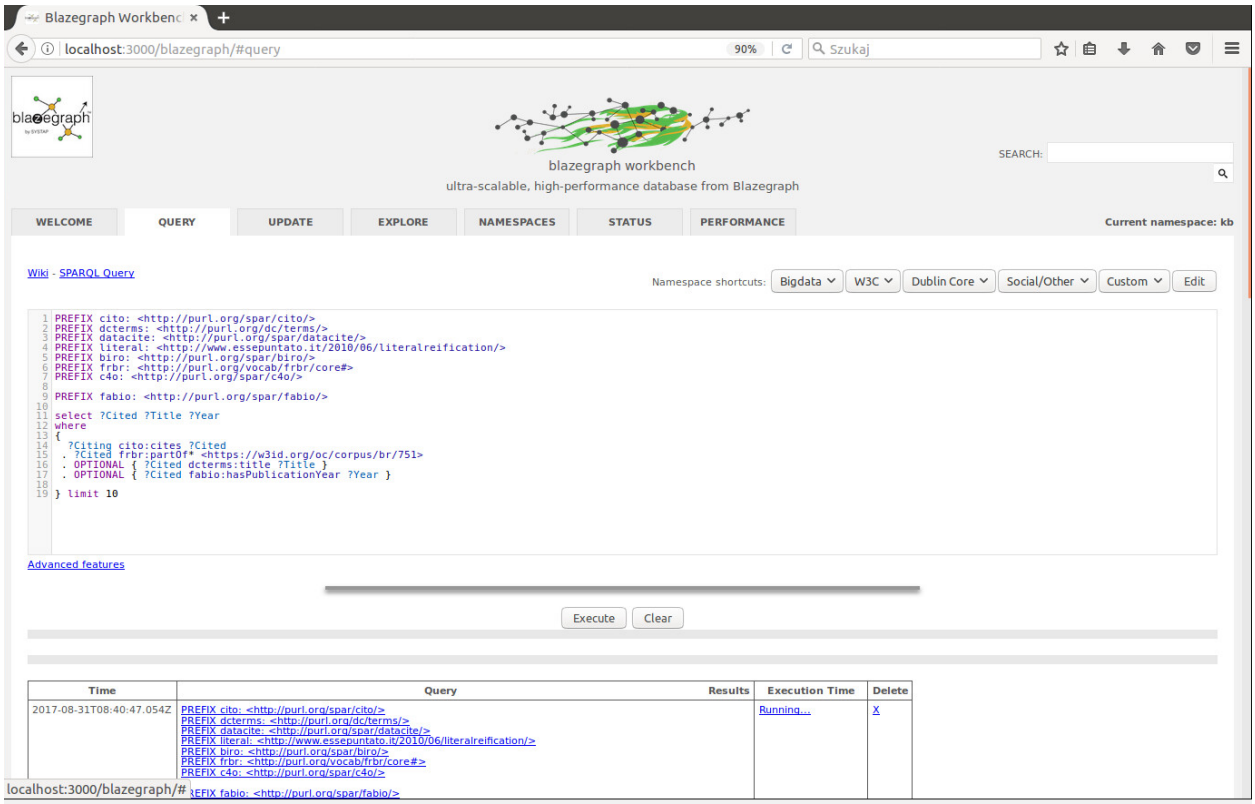
Rys. 4. Okno pobierania składowych korpusu OpenCitations
Źródło: Download. W: *OpenCitations* [online]. [Dostęp 09.09.2017]. Dostępny w: http://opencitations.net/corpus/br/1.html.

Konfiguracja i uruchomienie środowiska obliczeniowego

Do wdrożenia własnej instancji bazy danych niezbędny jest system operacyjny z zainstalowanym środowiskiem uruchomieniowym wirtualnej maszyny Javy. Pobrane archiwum ZIP zawiera szereg plików o rozszerzeniu DAR, będących z kolei składowymi innego systemu archiwizacji (Disk ARchive) umożliwiającego odtwarzanie pełnej struktury katalogów i plików o dużych rozmiarach. Po odtworzeniu struktury plików systemu OpenCitations z wykorzystaniem wymienionego programu na pierwszym poziomie drzewa katalogów znaleźć można wiele plików z rozszerzeniem „.sh", z których „run.sh" jest skryptem uruchomieniowym dla systemów rodziny Linux, umożliwiającym uruchomienie serwera BlazeGraph[5] wykorzystanego do obsługi bazy danych OpenCitations.

Komunikacja z systemem możliwa jest za pomocą usług sieciowych lub poprzez prostą aplikację WWW, która udostępniona jest domyślnie na porcie HTTP o numerze 3000. Można ją zatem uruchomić, wpisując w pole adresu przeglądarki internetowej http://localhost:3000/blazegraph/, natomiast okno z możliwością wysyłania zapytań do serwera znajduje się pod adresem http://localhost:3000/blazegraph/#query (rys. 5).

5 *Blazegraph* [online]. [Dostęp 9.09.2017]. Dostępny w: https://www.blazegraph.com/.



Rys. 5. Okno aplikacji do komunikacji z systemem BlazeGraph (zakładka Query)
Źródło: opracowanie własne.

Niestety, trzeba zauważyć, że możliwości formatowania czy pobierania wyników danego zapytania wyświetlonych w oknie przeglądarki są tutaj jeszcze bardziej ograniczone niż w aplikacji internetowej udostępnianej bezpośrednio ze stron twórców OCC. Z pomocą przychodzi jednak możliwość wykonywania zapytań i pobierania wyników w formatach CSV, XML oraz JSON z wykorzystaniem interfejsu usług sieciowych, z którym w najprostszym przypadku z poziomu systemów operacyjnych rodziny Linux komunikować się można komendą cURL. Producent serwera BlazeGraph objaśnia to szczegółowo w obszernej dokumentacji ilustrowanej bogato licznymi przykładami[6].

Przedstawione dotychczas informacje dają Czytelnikowi wystarczającą wiedzę o sposobie wdrożenia lokalnego środowiska obliczeniowego, możliwościach zadawania zapytań SPARQL oraz podglądu danych szczegółowych (zarówno w środowisku lokalnym, jak i udostępnionych jako aplikacja WWW) o jednostkach bibliograficznych i innych obiektach z nimi związanych, eksportowania wyników zapytań w zadanych formatach wprost z aplikacji WWW oraz uruchamiania zapytań i eksportowania ich wyników za pomocą komendy cURL w środowisku lokalnym. W dalszej części opracowania przedstawione zostaną przykładowe analizy bibliometryczne dotyczące cytowań artykułów publikowanych w ramach czasopisma PLOS ONE, choć nic nie stoi na przeszkodzie, aby na podstawie przedstawionych kroków realizować dalsze badania własne w odniesieniu do jakiejkolwiek innej, dowolnie wybranej, grupy prac naukowych.

[6] *Blazegraph – REST API* [online]. [Dostęp 9.09.2017]. Dostępny w: https://wiki.blazegraph.com/wiki/index.php/REST_API#QUERY.

Analizy realizowane bezpośrednio na bazie danych OCC z wykorzystaniem języka SPARQL

W ramach korpusu OCC gromadzone są przede wszystkim prace naukowe pochodzące z czasopism, tak więc spodziewać się należy znacznej przewagi liczebności artykułów tego typu wśród jednostek cytujących. Osoby analizujące relacje cytowań sprawdzić mogą, jak dużej ilości danych w ramach poszczególnych typów jednostek cytowanych można się spodziewać w całym korpusie. W tym celu można wykonać zapytanie:

```
PREFIX cito: <http://purl.org/spar/cito/>

select ?types (count ( ?types ) as ? counts)
{
        ?citing cito:cites ?cited
      . ?cited rdf:type ?types
}
group by ?types
order by desc ( ?counts )
```

Komenda PREFIX pozwala zdefiniować skrót cito dla ontologii opisanej pod adresem <http://purl.org/spar/cito/>. Definiowanie skrótów należy do dobrej praktyki (zwiększającej czytelność zapytań), zwłaszcza gdy pojęcia danej ontologii wykorzystywane są w zapytaniu wielokrotnie. W nawiasach klamrowych ujęto definicję podzbioru źródła obliczeń. Pierwsza trójka ograniczy wynik do wszystkich obiektów związanych relacją cytowania (czyli zwróci wszystkie podmioty i obiekty związane orzeczeniem cito:cites). Druga trójka spowoduje dodatkowo wyszukanie dla wcześniej znalezionych jednostek cytowanych obiektów związanych z nimi orzeczeniem rdf:type, czyli typów jednostek cytowanych. Typy te będą zagregowane i wyznaczone zostaną ich liczebności, a następnie wyświetlone w kolejności malejącej liczebności grup. Otrzymane wyniki przedstawiono poniżej.

Types	counts
<http://purl.org/spar/fabio/Expression>	8652350
<http://purl.org/spar/fabio/JournalArticle>	7270180
<http://purl.org/spar/fabio/BookChapter>	81829
<http://purl.org/spar/fabio/ProceedingsPaper>	27832
<http://purl.org/spar/fabio/Book>	17656
<http://purl.org/spar/fabio/ReferenceEntry>	16246
<http://purl.org/spar/fabio/DataFile>	6507
<http://purl.org/spar/fabio/ReportDocument>	2387
<http://purl.org/spar/fabio/Thesis>	741
<http://purl.org/spar/fabio/SpecificationDocument>	631
<http://purl.org/spar/fabio/Journal>	253
<http://purl.org/spar/fabio/Series>	193
<http://purl.org/spar/fabio/JournalIssue>	188
<http://purl.org/spar/fabio/ReferenceBook>	133
<http://purl.org/spar/fabio/ExpressionCollection>	51
<http://purl.org/spar/fabio/AcademicProceedings>	35
<http://purl.org/spar/fabio/BookSeries>	16

Należy zaznaczyć, że wszelkie analizy przedstawione w ramach niniejszego opracowania zrealizowane zostały dla danych udostępnionych w ramach stanu bazy danych z dnia 25 lipca 2017 r. i wyniki obliczeń na podstawie danych korpusu uaktualnianych w kolejnych miesiącach na pewno będą inne. Z uzyskanych danych wynika, że jednostki bibliograficzne inne niż artykuły z czasopism stanowią jedynie niewiele powyżej 2% całkowitej liczby cytowań. Pierwszy wiersz należy zignorować, gdyż dana jednostka może należeć do kilku klas (model z wielodziedziczeniem), a klasa wskazana przez pierwszy wiersz nie jest związana z formą wydawniczą. Oczywiście można by zmodyfikować zapytanie tak, by zwracane były wartości dotyczące jedynie form wydawniczych, jednak byłoby ono trudniejsze do opisania i dłużej by się wykonywało. Czas wykonania zapytania przedstawionego i tak już wynosił ponad 40 minut.

W kolejnym kroku sprawdzić można liczebność artykułów zgromadzonych w całym korpusie z podziałem na poszczególnych wydawców. W tym celu należy wykonać zapytanie:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>

SELECT  ?journaltitle ( count(*) as ?liczba )
{
        ?citing rdf:type <http://purl.org/spar/fabio/JournalArticle>
  .     ?citing frbr:partOf* ?container
  .     ?container dcterms:title ?journaltitle
  .     ?container rdf:type <http://purl.org/spar/fabio/Journal>
}
group by ?journaltitle
order by desc ( ?liczba )
```

Zapytanie to wyszukuje wszystkie jednostki, które są artykułami z czasopism. Czasopisma takie organizowane są w ramach „kontenerów" różnych typów („JournalIssue", „JournalVolume", „Journal") na kolejnych poziomach hierarchii. Zapytanie ogranicza przetwarzane trójki jedynie do tych, które związane są z wydawcą, dla którego jest wyszukiwany tytuł. Tytuły są następnie agregowane i wyświetlone w malejącej kolejności liczebności grup. Uzyskany wynik prezentuje bardzo obszerną listę (ponad 26 tysięcy), został więc ograniczony do pierwszych dziesięciu pozycji i przedstawiony poniżej.

journaltitle	Liczba
PLOS ONE - PLoS ONE	93056
Proceedings of the National Academy of Sciences	49679
Journal of Biological Chemistry	42100
Sci. Rep. – Scientific Reports	27150
Science	21621
Nature	20928
The Journal of Immunology	13327
Nucleic Acids Research	13182
Journal of Neuroscience	12557
Phys. Rev. Lett. – Physical Review Letters	12210

Z uzyskanych rezultatów wynika, że najwięcej artykułów w korpusie OCC zgromadzono dla czasopisma PLOS ONE.

Chcąc ograniczać całość danych opisujących cytowania do informacji związanych wyłącznie z PLOS ONE, warto poznać identyfikator (URI) tego wydawcy nadany w ramach korpusu tak, aby można się nim było posługiwać w ramach kolejnych zapytań. W tym celu wykonać można następujące zapytanie:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT * WHERE {
        ?citing rdf:type <http://purl.org/spar/fabio/Journal>
   .    ?citing dcterms:title ?title
   FILTER regex(?title, "^PLOS")
}
```

Komenda wyświetla rekord czasopisma o nazwie zaczynającej się od „PLOS". Pozwala to znaleźć jego identyfikator i przyjmuje on wartość: <https://w3id.org/oc/corpus/br/751>.

Konstrukcja języka SPARQL umożliwia prowadzenie dalszych badań, jak chociażby analizę liczby cytowań prac naukowych pochodzących z konkretnych zeszytów, artykułów konkretnych autorów, wpływu poszczególnych czasopism i wiele innych. Jednak prawdopodobnie nie dla każdego sposób ten będzie najbardziej intuicyjny i najszybciej prowadzący do zamierzonego celu. Dlatego podczas korzystania z innych narzędzi umożliwiających analizowanie struktur sieciowych może zajść potrzeba wyeksportowania interesującego w danym momencie fragmentu korpusu w postaci danych opisujących z osobna artykuły i oraz związki między nimi (ograniczone do relacji cytowań), gdyż taką właśnie formę danych akceptuje większość systemów analitycznych dedykowanych badaniom struktur sieciowych. Struktury takie w dalszej części artykułu nazywane są również grafami cytowań, lecz warto zauważyć, że do ich analizowania nie jest potrzebna znajomość całości dziedziny teorii grafów, a jedynie jej niewielki wycinek. Więcej o możliwości zastosowania tych struktur w dziedzinie bibliometrii i webometrii znaleźć można w osobnym opracowaniu autorki[7].

Analizy realizowane w narzędziu Gephi

Jako przykładowe narzędzie dalszych analiz użyta zostanie aplikacja Gephi, która mimo że ciągle dostępna jest jedynie w fazie rozwojowej (tzw. wersja beta), istnieje już od ponad ośmiu lat i, jako narzędzie przyjazne i o łatwo rozszerzalnych możliwościach za pomocą bogatej biblioteki komponentów (wtyczek – ang. plugins), wybierana jest chętnie przez wielu badaczy.

Podstawowym formatem pliku składowania informacji o strukturach sieciowych jest GEXF oparty na XML. Jego zastosowanie daje wiele korzyści, co przedstawiono w jednym[8]

[7] KAMIŃSKA, A.M. *Zastosowanie struktur grafowych do analiz bibliometrycznych i webometrycznych. Modele i metody* (w druku).
[8] KAMIŃSKA, A.M. Od druków źródłowych po mapy nauki. Bibliograficzna baza danych GRUBA. W: KOWALSKA, M., OSIŃSKA, V. (red.). *Wizualizacja informacji w humanistyce.* Toruń: Wydaw. Uniwersytetu Mikołaja Kopernika, 2017 (w druku).

z wcześniejszych opracowań autorki, jednak z bazy danych BlazeGraph o wiele łatwiej (tzn. jedynie z wykorzystaniem zapytań SPARQL) dane będzie wyeksportować w postaci plików CSV. Dla pliku opisującego połączenia sieci (w naszym przypadku relacje cytowania) aplikacja Gephi oczekuje istnienia przynajmniej dwóch kolumn o nazwach „Source" i „Target" zawierających odpowiednio identyfikatory obiektu źródłowego i docelowego (w naszym przypadku artykułów cytujących i cytowanych). Wynik poniższego zapytania SPARQL zapisany do pliku stanowić może bezpośrednio źródło informacji opisujące cytowania, którym zasilić możemy aplikację Gephi:

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>

SELECT
(
replace(str(?Citing),'https://w3id.org/oc/corpus/br/','') as ?Source
)
(
replace(str(?Cited),'https://w3id.org/oc/corpus/br/','') as ?Target
)
WHERE
{
     ?Citing cito:cites ?Cited
   . ?Cited frbr:partOf* <https://w3id.org/oc/corpus/br/751>
}
```

Zapytanie przeszukuje wszystkie cytowania i ogranicza wyniki do rekordów, dla których jednostka cytowana zawiera się w zeszytach czasopisma o wcześniej znalezionym identyfikatorze wskazującym na PLOS ONE. We frazie SELECT dodatkowo zastosowano funkcje usuwające prefiksy charakterystyczne dla identyfikatorów URI, pozostawiając jedynie wartość liczbową, co pozwoli na otrzymanie bardziej przejrzystej postaci identyfikatorów.

Zapisując wynik zapytania w formacie CSV, otrzymujemy plik z nagłówkami o wymaganych nazwach „Source" i „Target", dzięki czemu możemy zaimportować go funkcją „Import spreadsheet" jako plik krawędzi (ang. *edges*) do aplikacji Gephi z poziomu zakładki „Data laboratory". Zaznaczając opcję „create missing nodes" możemy wczytać plik, a system automatycznie wygeneruje wierzchołki (ang. *nodes*) na podstawie identyfikatorów znalezionych w kolumnach „Source" i „Target". Pozwoli to już co prawda na analizę struktury sieciowej, jednak tak wygenerowane wierzchołki reprezentujące artykuły nie będą zawierać żadnych informacji (poza identyfikatorami pozwalającymi na identyfikację jednostek bibliograficznych w OCC) je opisujących. Chcąc móc obserwować podstawowe informacje o artykułach w systemie Gephi, należy samodzielnie stworzyć plik opisujący wierzchołki wraz z dodatkowymi informacjami, które je opiszą (tytuł artykułu, rok wydania itp.). Przyjmując, że prowadzone analizy skupiają się jedynie na artykułach z czasopisma PLOS ONE, wystarczy wygenerować plik opisujący jedynie te jednostki bibliograficzne. Możemy to uczynić następującą komendą:

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX fabio: <http://purl.org/spar/fabio/>

SELECT
```

```
distinct
( replace(str(?Cited),'https://w3id.org/oc/corpus/br/','') as ?Id )
( ?Title as ?Label )
?Year
WHERE
{
    ?Citing cito:cites ?Cited
  . ?Cited frbr:partOf* <https://w3id.org/oc/corpus/br/751>
  . OPTIONAL { ?Cited dcterms:title ?Title }
  . OPTIONAL { ?Cited fabio:hasPublicationYear ?Year }
}
```

Powyższe zapytanie zwraca wszystkie unikatowe identyfikatory jednostek cytowanych i publikowanych w czasopiśmie PLOS ONE. Dodatkowo, jeśli będą one opisane tytułami i latami wydań, informacje te również będą zawarte w odpowiedzi. Warto zwrócić uwagę, że wynik zapisany w formacie CSV identyfikatory będzie opisywał nagłówkiem o nazwie „Id", a tytuły nagłówkiem o nazwie „Label". Są to nazwy oczekiwane przez system Gephi. Natomiast nagłówek o nazwie „Year" opisujący kolumnę z latami publikacji stanowić będzie atrybut dodatkowo opisujący wierzchołek.

Kolejność postępowania tworzenia grafu cytowań w systemie Gephi przy pomocy dwóch wygenerowanych powyżej plików jest więc następująca:

1. Utworzyć nowy projekt („New project") w systemie Gephi;
2. Zaimportować plik jednostek cytowanych (opcja „Nodes table") z opisami tytułów i lat publikacji;
3. Zaimportować plik krawędzi cytowań (opcja „Edges table") z użyciem opcji tworzenia brakujących wierzchołków („Create missing node").

Tak zasilony system gotowy jest już do rozpoczęcia analiz. Więcej o możliwościach importowania danych w formacie CSV znaleźć można na stronach[9] twórców aplikacji.

Dla danych wczytanych w powyżej opisany sposób obliczono liczbę cytowań dla poszczególnych artykułów jako stopień wchodzący wierzchołka (ang. *in-degree*). Zostały więc uwzględnione wszystkie jednostki cytujące (pochodzące z PLOS ONE jak i wszystkie inne). Jako że przykładowym celem analiz są artykuły publikowane w PLOS ONE, graf ograniczono tylko do takich jednostek. Uzyskano więc graf cytowań pomiędzy jednostkami PLOS ONE, ale zawierający informacje o liczbie wszystkich cytowań przypadających na dany artykuł.

Wartości liczby cytowań przedstawione zostały na rys. 6 z którego wynika, że najczęściej cytowanym artykułem (200 razy) jest „Fast tree…", zaś kolejne („REVIGO Summarizes…", „Leishmaniasis Worldwide..." i kolejne) dzieli już od lidera spora różnica cytowań. Ich liczba uwzględniona jako wielkość wierzchołka z wykorzystaniem algorytmu rozmieszczania bazującym na symulacji sił grawitacji (ang. *atlas force*) pozwoliły na uzyskanie mapy (rys. 7).

[9] *Gephi makes graphs handy – CSV format* [online]. [Dostęp 9.09.2017]. Dostępny w: https://gephi.org/users/supported-graph-formats/csv-format/.

Overview | Data Laboratory | Preview

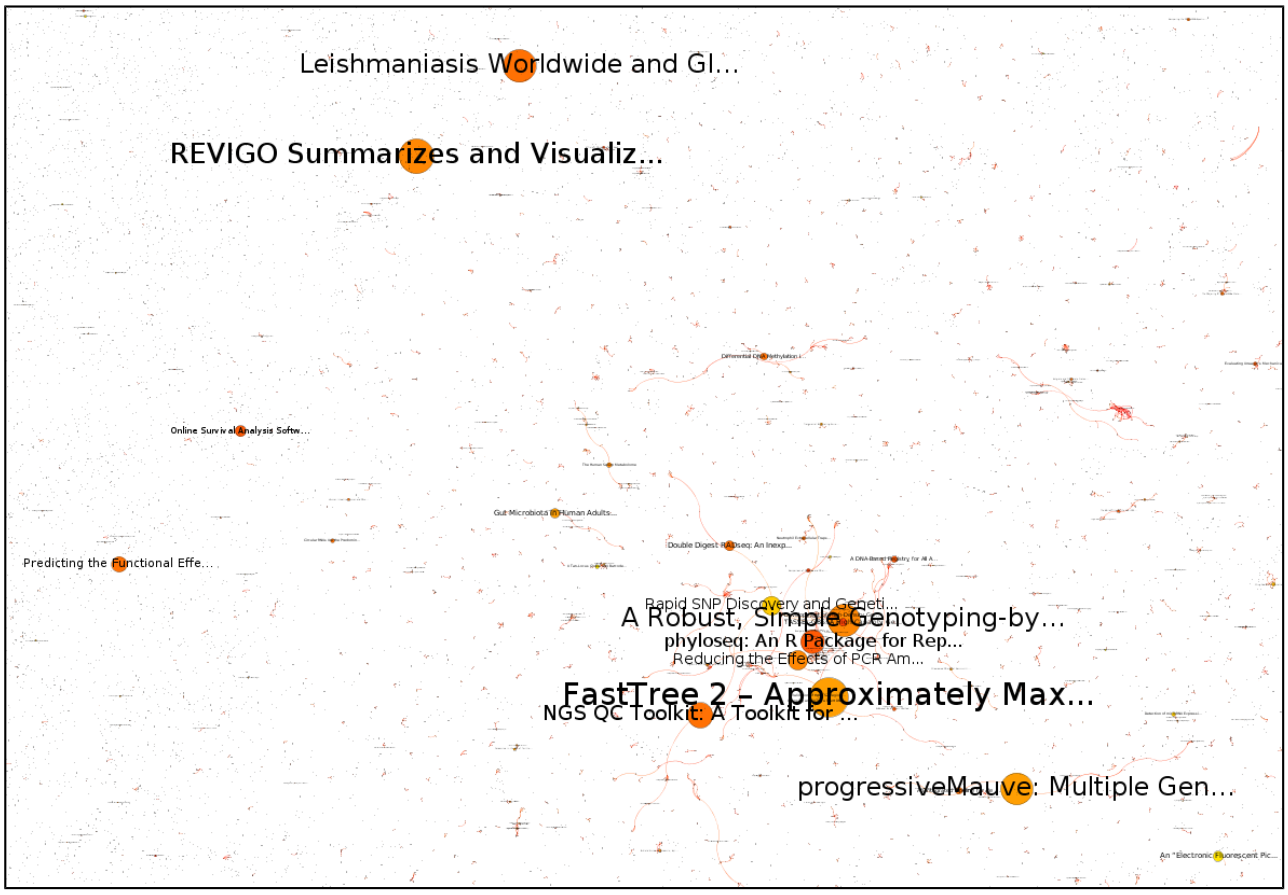
Workspace 1

Data Table

Nodes | Edges | Configuration | Add node | Add edge | Search/Replace | Import Spreadsheet | Export table | More actions | Filter: | Id

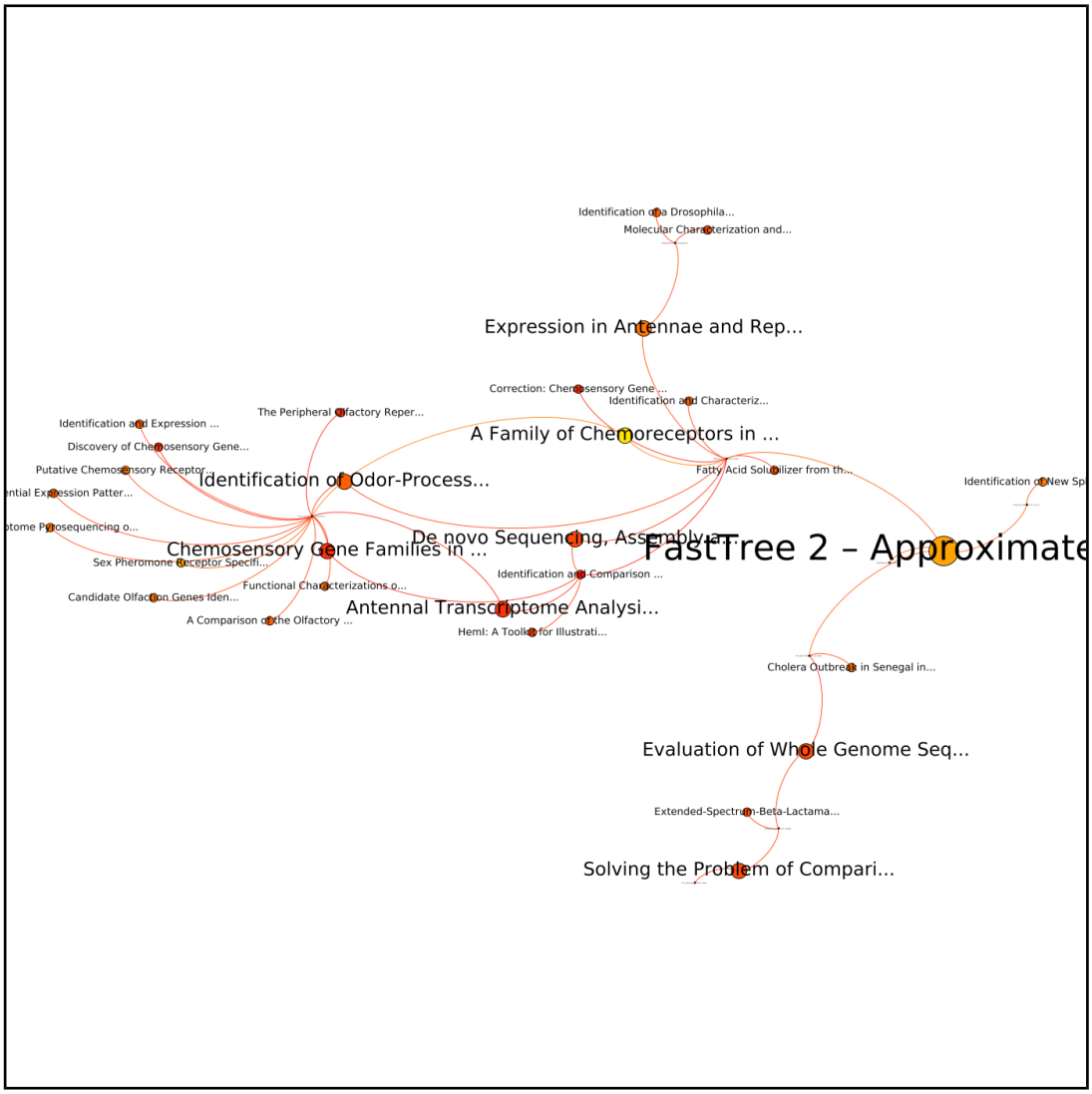
Id	Label	Interval	Year	In-D...	Out-De...	Degree
197200	FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments		2010	200	0	200
338150	REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms		2011	136	0	136
272393	Leishmaniasis Worldwide and Global Estimates of Its Incidence		2012	124	0	124
67253	A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species		2011	122	0	122
172916	progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement		2010	120	0	120
90095	NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data		2012	92	0	92
92325	phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data		2013	86	0	86
193476	Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies		2011	73	0	73
172409	Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers		2008	72	0	72
231187	REST: A Toolkit for Resting-State Functional Magnetic Resonance Imaging Data Processing		2011	69	0	69
148079	Predicting the Functional Effect of Amino Acid Substitutions and Indels		2012	62	0	62
150983	Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in No...		2013	50	0	50
372457	BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics		2013	50	0	50
360427	Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-...		2012	47	0	47
770765	An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets		2007	47	0	47
67400	Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation		2010	46	0	46
291850	Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults		2010	45	0	45
129693	TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline		2014	41	0	41
198553	Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-S...		2012	39	0	39
7847	An Integrated Pipeline for de Novo Assembly of Microbial Genomes		2012	38	0	38
350669	Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement		2014	38	0	38
91967	Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Diseas...		2012	38	0	38
708381	Generation of Breast Cancer Stem Cells through Epithelial-Mesenchymal Transition		2008	37	0	37
690367	A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System		2013	36	0	36
154876	Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology		2012	35	0	35
271272	The Human Serum Metabolome		2011	31	0	31
815586	Computer Therapy for the Anxiety and Depressive Disorders Is Effective, Acceptable and Practical Health Care:...		2010	30	0	30
1119446	A New Mesenchymal Stem Cell (MSC) Paradigm: Polarization into a Pro-Inflammatory MSC1 or an Immunosuppr...		2010	30	0	30
176407	A One Pot, One Step, Precision Cloning Method with High Throughput Capability		2008	30	0	30
741297	Serum MicroRNAs Are Promising Novel Biomarkers		2008	30	0	30

Rys. 6. Jednostki PLOS ONE w malejącej kolejności liczby cytowań
Źródło: opracowanie własne.



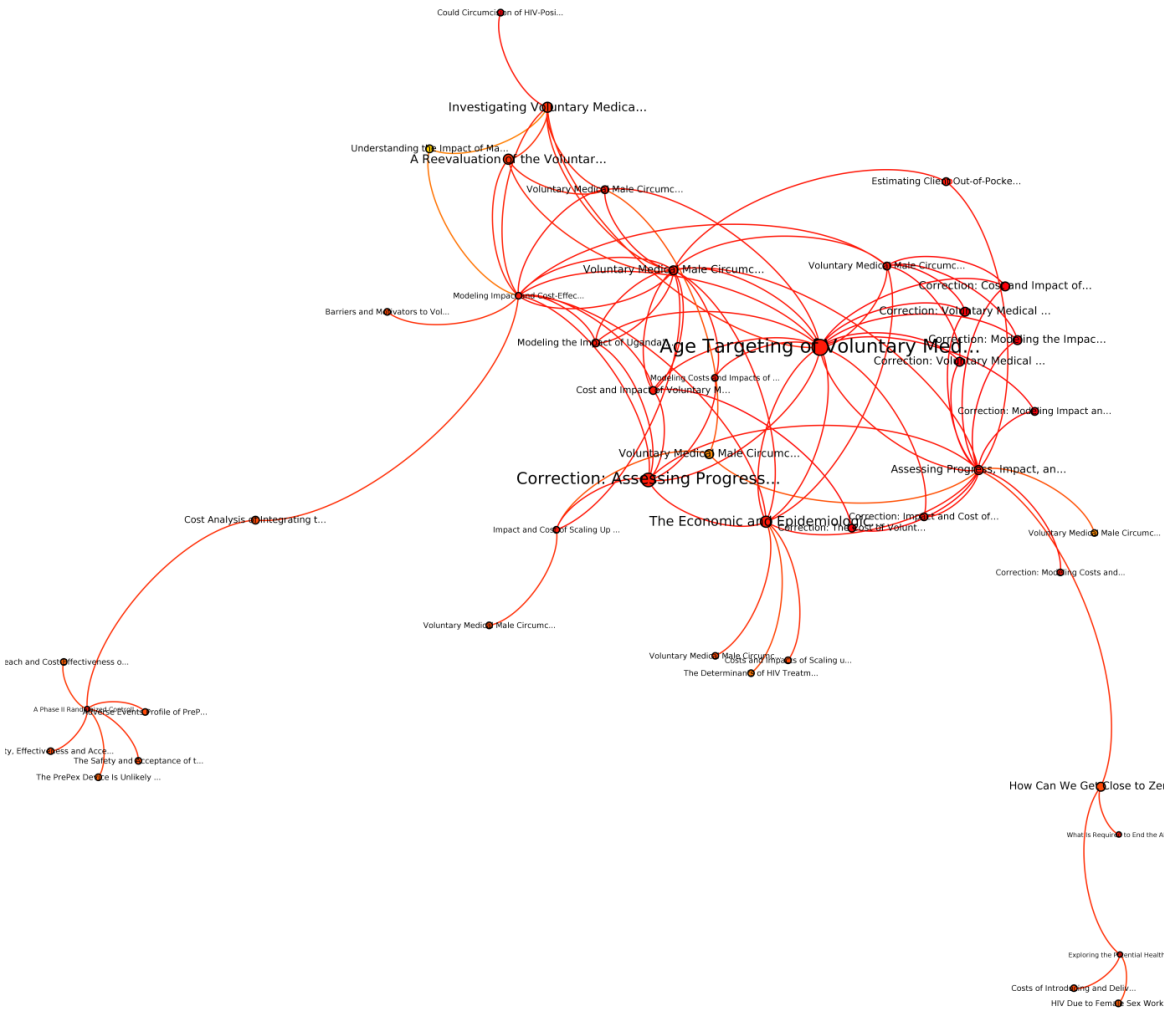
Rys. 7. Mapa cytowań artykułów z PLOS ONE
Źródło: opracowanie własne.

Mapa ta pozwala odkryć zależności nieobserwowalne w formie tabelarycznej. Łuki łączące poszczególne wierzchołki obrazują relację cytowania, która nie jest relacją zwrotną. Jest to więc graf skierowany, a kierunek cytowania zgodny jest z kierunkiem ruchu wskazówek zegara danego łuku. Widać tutaj, że artykuły „REVIGO Summarizes…" oraz „Leishmaniasis Worldwide..." cytowane są często, ale przez jednostki pochodzące spoza czasopisma PLOS ONE. Natomiast powiązana ze sobą grupa artykułów „Fast Tree..", „A Robust…" i innych wskazuje na możliwość ich wzajemnych związków tematycznych. Warto zwrócić uwagę, że jednostki o bardzo małej liczbie cytowań mają również bardzo małe wierzchołki i ich etykiety. Stanowią one jedynie mniej istotne tło analiz. Choć przedstawiając „mapę" jako statyczny obraz, trudno dostrzec nazwy najmniejszych z nich, to oczywiście prowadząc analizy w narzędziu Gephi, możliwe jest ich interaktywne przybliżanie i skupianie się na wybranych podobszarach mapy. Na podstawie tak stworzonej mapy można budować hipotezy, których weryfikację może ułatwić bliższe przyjrzenie się poszczególnym jednostkom. Dla jednostki „Fast Tree…" przedstawiono na rys. 8 graf artykułów, z którymi jest związany. Widzimy tutaj, że mimo wysokiej pozycji w rankingu artykuł ten cytowany jest jedynie poprzez trzy inne jednostki publikowane w PLOS ONE.



Rys. 8. Graf artykułów związanych z „Fast Tree…"
Źródło: opracowanie własne.

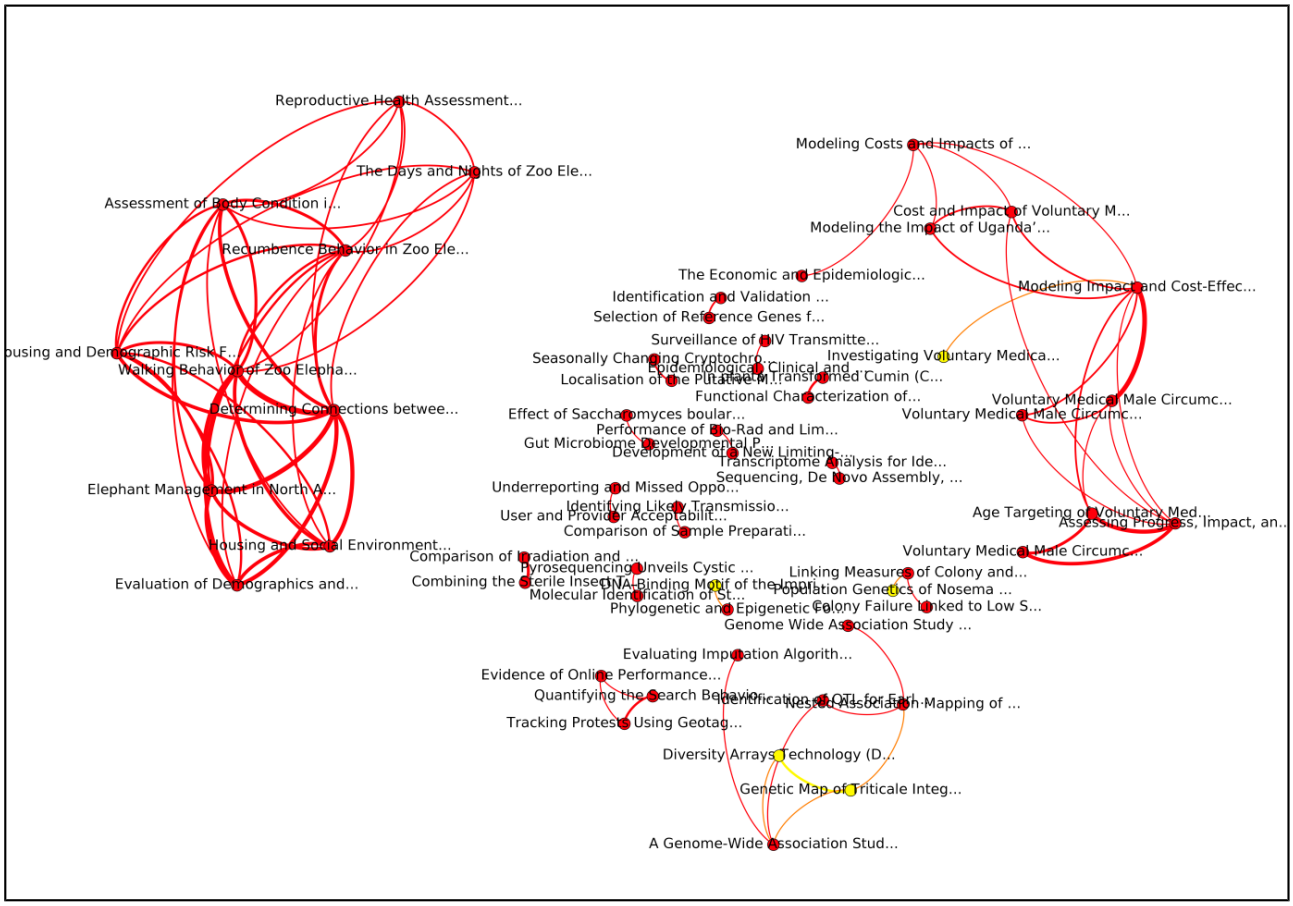
Zupełnie inaczej ma się sytuacja w przypadku artykułu „Age Targeting…" dla którego graf przedstawiono na rys. 9. Mimo że jednostka ta uzyskała „jedynie" dziewięć cytowań, to bogata sieć powiązań (zarówno jednostek go cytujących, jak i takich, które cytuje) świadczyć może o dużym zainteresowaniu tematyką, poruszaną przez autorów publikujących w PLOS ONE.



Rys. 9. Graf artykułów związanych z „Age Targeting..."
Źródło: opracowanie własne.

O ile mapy cytowań stanowiące grafy skierowane obrazujące bezpośrednie i jednostkowe fakty cytowania pomiędzy pracami naukowymi mogą obrazować dość długie ścieżki cytowań trudne do obrazowania i analizowania bez użycia takich środków jak interaktywne przybliżanie czy filtrowanie grafów, o tyle zastosowanie miary współcytowań pozwala otrzymać mapy łatwiejsze do analizowania jako obrazy statyczne. Miara ta dla dwóch wybranych jednostek bibliograficznych przyjmuje wartość równą liczbie dokumentów, które cytują te jednostki równocześnie. Otrzymany graf jest więc tym razem grafem nieskierowanym, ale za to ważonym – co oznacza, że łuki łączące wierzchołki (krawędzie) nie mają kierunku, ale posiadają swoją charakterystyczną liczbę (wagę), która w celach lepszej wi-

zualizacji może wpływać na jego grubość. Przykład wizualizacji tej miary dla jednostek analizowanego korpusu (podzbiór dokumentów PLOS ONE pochodzących z OCC) obrazuje rys. 10, który czytać można w ten sposób, że dokumenty połączone łukiem są ze sobą związane tym bardziej, im łuk ten jest grubszy.



Rys. 10. Mapa współcytowanych artykułów
Źródło: opracowanie własne.

Opisane powyżej analizy wykonane z użyciem narzędzia Gephi przedstawiają jedynie podstawy jego zastosowań do badań bibliometrycznych. Oprócz cytowań pomiędzy jednostkami bibliograficznymi możliwe są również analizy na poziomie większej agregacji (np. pomiędzy czasopismami czy instytucjami), analizy współpracy pomiędzy badaczami (zarówno w sensie cytowań czy relacji współautorstwa), wizualizacje wskaźników bibliometrycznych, takich jak liczba cytowań, miara powiązań bibliograficznych czy miara współcytowań pomiędzy jednostkami bibliograficznymi. Każde z tych zagadnień na przykładzie danych pochodzących z krajowej bibliograficznej bazy CYTBIN przedstawione zostało we wcześniejszym opracowaniu autorki[10], a dodatkowo w opracowaniu[11] opublikowanym w ramach materiałów pokonferencyjnych konferencji „Wizualizacja Informacji w Humanistyce" (23–24 marca 2017 r.). Warto dodać, że platforma Gephi umożliwia

[10] KAMIŃSKA, A.M. Wizualizacje wybranych wskaźników bibliometrycznych na przykładzie bibliograficznej bazy danych CYTBIN. *Toruńskie Studia Bibliologiczne* 2017, 2 (19) (w druku).
[11] KAMIŃSKA, A.M. Od druków źródłowych po mapy nauki. Bibliograficzna baza danych GRUBA. W: KOWALSKA, M., OSIŃSKA, V. (red.), dz. cyt.

również obliczanie wielu miar stosowanych w zagadnieniach analizy sieci społecznościowych. Propozycje wykorzystania tych miar na gruncie bibliometrii i innych badań nad rozwojem nauki przedstawione zostały w kolejnym opracowaniu[12].

Wnioski

Artykuł, przedstawiając przykładowe analizy na danych pochodzących z otwartego korpusu cytowań, pokazuje w formie studium przypadku możliwość ekstrakcji podzbioru danych ze wspomnianego korpusu w formie plików formatu CSV, którymi opisać można grafy cytowań. Ukazano również możliwość prowadzenia analiz bibliometrycznych w narzędziu dedykowanym analizom struktur sieciowych, co rozszerza potencjał analiz o możliwości stawiania hipotez trudnych do dostrzeżenia w danych zgromadzonych w tradycyjnych układach tabelarycznych.

Publikowanie danych o cytowaniach prac naukowych w formie powszechnego dostępu otwiera nowe możliwości analiz rozwoju dziedzin nauki. Badacze nie musząc już ograniczać się do limitowanego dostępu do komercyjnych baz danych czy rejestrowania danych bibliograficznych z autopsji[13], zyskują możliwość stosunkowo łatwego pozyskania wiarygodnych danych bibliograficznych.

Przedstawione przykłady analiz ukierunkowane są na cel dydaktyczny i nie dają podstaw do wyciągania prawomocnych wniosków co do znaczenia poszczególnych artykułów dla rozwoju nauki. Im większy będzie zasięg korpusu OCC i im dłuższy stanie się jego retrospektywny horyzont czasowy, tym bardziej wiarygodne będą zaobserwowane zależności. Choć z jednej strony według obiegowych opinii cykl życia prac nauk technicznych jest stosunkowo krótki, to jednak wydłużony czas ich publikowania w tradycyjnym modelu powoduje spore opóźnienia. Rozwój koncepcji otwartego dostępu sprzyjać będzie niewątpliwie zarówno skróceniu tej bezwładności, jak i możliwości szybszej obserwacji zmian zachodzących w rozwoju gałęzi poszczególnych dziedzin nauki.

Bibliografia:

1. *Blazegraph – REST API* [online]. [Dostęp 9.09.2017]. Dostępny w: https://wiki.blazegraph.com/wiki/index.php/REST_API#QUERY.
2. *Blazegraph* [online]. [Dostęp 9.09.2017]. Dostępny w: https://www.blazegraph.com/.
3. *Gephi makes graphs handy – CSV format* [online]. [Dostęp 9.09.2017]. Dostępny w: https://gephi.org/users/supported-graph-formats/csv-format/
4. KAMIŃSKA, A.M. Od druków źródłowych po mapy nauki. Bibliograficzna baza danych GRUBA. W: KOWALSKA, M., OSIŃSKA, V. (red.). *Wizualizacja informacji w humanistyce*. Toruń: Wydaw. Uniwersytetu Mikołaja Kopernika, 2017.
5. KAMIŃSKA, A.M. OpenCitations – otwarty indeks cytowań publikacji naukowych. *Biuletyn EBiB* [online]. 2017, No 176. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/551.

[12] KAMIŃSKA, A.M. Zastosowanie metod analizy sieci społecznościowych w bibliometrii i webometrii. Miary i narzędzia. *Nowa Biblioteka. Usługi, technologie informacyjne i media* 2018, 2 (29) (w druku).
[13] KAMIŃSKA, A.M. Tam, gdzie zaczyna się bibliometria, czyli jak pozyskać materiał analityczny z autopsji. *Biuletyn EBiB* [online]. 2017, No 173. [Dostęp 9.09.2017]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/534.

6. KAMIŃSKA, A.M. Tam, gdzie zaczyna się bibliometria, czyli jak pozyskać materiał analityczny z autopsji. *Biuletyn EBiB* [online]. 2017, No 173. [Dostęp 16.08.2017]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/534.
7. KAMIŃSKA, A M. Wizualizacje wybranych wskaźników bibliometrycznych na przykładzie bibliograficznej bazy danych CYTBIN. *Toruńskie Studia Bibliologiczne* 2017, 2 (19) (w druku).
8. KAMIŃSKA, A.M. Zastosowanie metod analizy sieci społecznościowych w bibliometrii i webometrii. Miary i narzędzia. *Nowa Biblioteka. Usługi, technologie informacyjne i media* 2018, 2 (29) (w druku).
9. KAMIŃSKA, A.M. Zastosowanie struktur grafowych do analiz bibliometrycznych i webometrycznych. Modele i metody (w druku).
10. PERONI, S., DUTTON, A., GRAY, T., SHOTTON, D. Setting our bibliographic references free: towards open citation data. *Journal of Documentation* [online]. 2015, 71 (2), s. 253–277. [Dostęp 9.09.2017]. Dostępny w: http://speroni.web.cs.unibo.it/publications/peroni-2015-setting-bibliographic-references.pdf.
11. *RDF* [online]. W3C, 2017. [Dostęp 9.09.2017]. Dostępny w: https://www.w3.org/RDF/.
12. *SPARQL Query Language for RDF* [online]. W3C, 2017. [Dostęp 9.09.2017]. Dostępny w: https://www.w3.org/TR/rdf-sparql-query/.