Veslava Osińska
Uniwersytet Mikołaja Kopernika
wieo@umk.pl

Gephi – nowoczesne narzędzie do wizualizacji danych bibliometrycznych

Streszczenie: Po warsztatach, które odbyły się podczas konferencji „Wizualizacja informacji w humanistyce” autorka opracowała instrukcję wykorzystania aplikacji Gephi do przygotowania wizualizacji danych bibliometrycznych. W bibliotekach naukowych powstają sekcje bibliometryczne, które przygotowują dla administracji uczelni informacje o publikacjach naukowych pracowników macierzystej instytucji i ich cytowaniach. Informacje te lepiej prezentują się i przemawiają do administratorów, jeśli są przejrzyście zaprezentowane. Niniejszy artykuł ma zachęcić osoby odpowiedzialne za przygotowywanie informacji bazujących na danych bibliometrycznych, w szczególności bibliotekarzy, do korzystania z Gephi i ułatwić wykorzystanie różnych funkcji aplikacji. Może także pomóc bibliotekarzom w zdobywaniu nowych kwalifikacji i umiejętności.

Słowa kluczowe: wizualizacja danych, bibliometria, naukometria, szkolenie bibliotekarzy, instrukcja

Abstract: Based on the experience gained during the workshop that accompanied a conference "Visualization of information in humanities", the author prepared an instruction (tutorial) of the use of Gephi application for the visualization of bibliometric data. In research libraries bibliometric sections prepare information on the scientific output of their parent institutions. That information is more appealing to research authorities including administration officers if it is clearly visualized. The aim of this paper and the instruction is to encourage specialists responsible for the presentation of bibliometric data, especially librarians, to use various functionalities of Gephi in their work. It may also help librarians gain new skills and qualifications.

Keywords: data visualisation, bibliometrics, scientometrics, training for librarians, tutorial

Podczas konferencji „Wizualizacja informacji w humanistyce” na Uniwersytecie Mikołaja Kopernika w Toruniu w dniach 24-25 marca 2017 r., adresowanej przede wszystkim do rodzimych bibliologów i informatologów, okazało się, że coraz częściej sięgają oni po graficzne formy prezentacji informacji i wiedzy. Dlatego warsztaty, umożliwiające zapoznanie się z różnymi technikami i narzędziami wizualizacji informacji, były bardzo przydatne. Bibliometrzy i naukometrzy dowiedzieli się, jak wspomagać analizę dużych zbiorów danych zaawansowaną wizualizacją.

Automatyczne tworzenie raportów dla danych z baz Web of Science (WoS) zapewniają rozwiązania komercyjne, np. demonstrowany na wspomnianej konferencji program InCites oferowany przez Clarivate Analytics. Istnieją również inne rozwiązania, pozwalające na pracę analityków z zestawem darmowych aplikacji. Poznanie tych bardzo zróżnicowanych narzędzi oraz formatów plików przez nie obsługiwanych okazało się niezwykle istotne dla uczestników warsztatów. Z uwagi na możliwość udziału w warsztatach niewleikiej grupy specjalistów, autorka opracowała niniejszy instruktaż (tutorial), który ma służyć między innymi popularyzacji narzędzi do wizualizacji danych w kręgach bibliometrów i bibliotekarzy.

Na warsztatach w Toruniu ćwiczono z dwoma aplikacjami: Pajek i Gephi. Obie przeznaczone są do modelowania i analizy sieci społecznościowych (ang. *social network analysis*, SNA) z zastosowaniem popularnej obecnie metody badania struktur społecznych, której korzenie wywodzą się z nauki o sieci (ang. *network science*). W bibliometrii duże zbiory danych, dotyczących cytowań, współcytowań czy współautorstwa, odwzorowują sieciowe powiązania i zależności (w odróżnieniu od hierarchii), co motywuje badaczy do wykorzystania metod SNA w przetwarzaniu i modelowaniu takich sieci.

Format NET

Pajek jest oprogramowaniem stworzonym w połowie lat 90. przez słoweńskich naukowców Andreja Mrvara i Vladimira Batagelja. Ma bardzo rozbudowane funkcje, lecz nie jest przyjazne użytkownikowi, a to obecnie jest kluczowa kwestia. Do zastosowań bibliometrycznych służy aplikacja Bibexcel autorstwa znanego szwedzkiego bibliometry Olle Persona, która ma za zadanie przygotowywanie danych pobranych z WoS lub Scopusa do formatu wejściowego Pajeka – NET. Niestety, aplikacja Bibexcel, pomimo licznych opcji, ma wyjątkowo przestarzały interfejs, co zniechęca współczesnych użytkowników do stosowania jej w analizach bibliometrycznych. Z gradem krytyki tego programu autorka miała do czynienia podczas ćwiczeń ze studentami: dla młodych ludzi taka aplikacja jest nie do zaakceptowania. Dlatego poniżej zostanie zademonstrowane alternatywne narzędzie.

Format NET (własny format Pajeka) jest obecnie ogólnym standardem pliku tekstowego, zawierającego informację o strukturze sieci. Wszystkie aplikacje do SNA są kompatybilne z tym formatem. Ma on niezwykle intuicyjną składnię. Sieć składa się z węzłów (ang. *nodes*) i połączeń pomiędzy nimi – krawędzi (ang. *edges*). Każdy format sieciowy powinien zawierać podstawowe informacje o węzłach i krawędziach. Format NET opisuje węzły i krawędzie w osobnych sekcjach *Vertices* i *arcs:*
Przykład:

```
*Vertices 110

1 "nazwa węzła 1"
2 "nazwa węzła 2"
3 "nazwa węzła 3"
4 "nazwa węzła 4"
5 "nazwa węzła 5"
*arcs

1 2 34
2 3 56
4 1 9
1 3 20
```

Sekcja *Vertices* w danym przypadku opisuje 110 obiektów (węzłów). Składa się z dwóch kolumn: identyfikatora węzła oraz jego nazwy. W sekcji *arcs* kolumny pierwsza i druga podają identyfikatory połączonych węzłów, w trzeciej zapisywana jest waga, czyli liczbowa miara połączenia. Im większa, tym silniejsza relacja pomiędzy węzłami i tym bliżej siebie będą one się znajdowały w sieci. W analizie cytowań za wagi posłuży liczba wspólnych cytowań w rozważanych artykułach (bibliografia łączona) albo liczba prac cytujących dwa dokumenty (współcytowania). W analizie współautorstwa wagą będzie liczba wspólnych prac. W takim formacie przygotowanie własnych danych nie powinno stanowić problemu, o ile się umie operować kolumnami liczb w Excelu oraz eksportować je do formatu CSV, który kolumny oddziela tabulatorami lub średnikami.

Dane ze Scopusa

Od 2010 r. zaczęły pojawiać się programy do SNA nowej generacji. Jednym z nich jest Gephi, środowisko do wizualizacji, które szybko zdobyło uznanie użytkowników: zarówno informatyków, jak i biologów (pracujących z masowymi danymi dotyczącymi kodowania białek) oraz przedstawicieli nauk społecznych. Z Gephi jest związana liczna społeczność in-

formatyków, którzy rozwijają funkcje i dodatki do programu. Bibliometrzy powinni znać to narzędzie ze względu na niezwykle intuicyjny interfejs, możliwości statystyczne oraz generowanie pięknych wizualizacji.

W niniejszym tutorialu zostały wykorzystane dane z bazy Scopus do stworzenia sieci cytowań osoby znanej w środowisku polskich bibliotekarzy – mgr Bożeny Bednarek-Michalskiej. Mały rozmiar próby – 2 artykuły oraz 4 cytowania jest korzystny dla przeprowadzenia szczegółowej analizy kroków procedury mapowania (wizualizacji) i przyjrzenia się różnym opcjom.

Po odfiltrowaniu w bazie Scopus dokumentów według nazwiska autora otrzymujemy listę publikacji. Zaznaczenie opcji *All* umożliwia eksport wszystkich wyników (opcja *Export*) po uprzednim wybraniu w oknie dialogowym interesujących nas metadanych. Dla formatu pliku wyjściowego należy zaznaczyć CSV. Będzie to plik zawierający opisy publikacji danego autora, nazwijmy go „publikacje.csv". Dla zaznaczonej listy dostępna jest opcja *View cited by*, która wyświetli wszystkie cytowania z bazy Scopus. Wyniki w identyczny sposób zapiszemy w formacie CSV w pliku o nazwie np. „cytowania.csv".

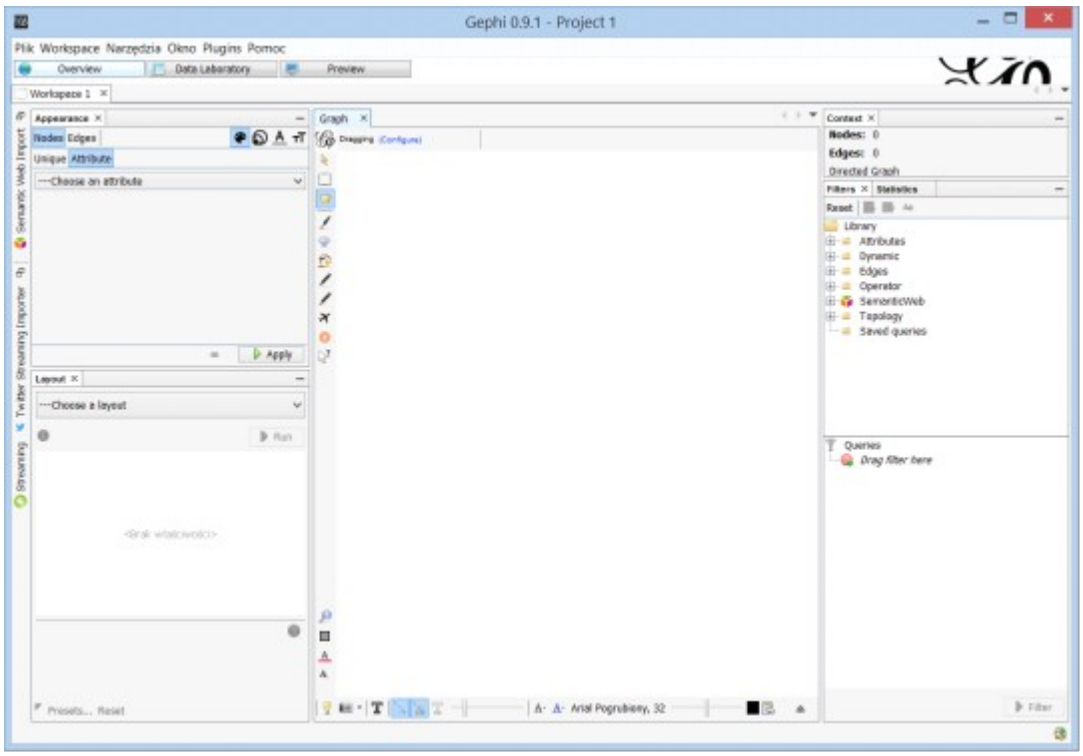
Praca z aplikacją Gephi – instrukcja

Plik instalacyjny Gephi można pobrać ze strony portalu Gephi (https://gephi.org/[1]). Obecnie (lipiec 2017 r.) dostępna jest w wersja 0.9.1. Aplikacja wymaga zainstalowania aktualnego oprogramowania Javy. Po zainstalowaniu i otwarciu programu, w oknie dialogowym trzeba wybrać *New Project*, ponieważ w nowym projekcie zaczniemy operować na danych, które zaimportujemy z plików CSV. Innego wyboru dokonujemy w przypadku plików formatów sieciowych, takich jak np. NET, GEXF, GDF – wybieramy wówczas opcję *Open Network File*.

Aplikacja Gephi ma poza angielską wersją dostępne inne wersje językowe. Można je wybrać w menu przez opcję *Narzędzia>Language*. Liczba języków do wyboru świadczy o licznej społeczności informatyków i tłumaczy, którzy zajmują się tłumaczeniem interfejsu. Jeśli chodzi o wersję polską, to nie wszystko jest przetłumaczone, co pokazuje, że w społeczności brakuje polskich wolontariuszy.

W obszarze roboczym pracuje się w trzech podstawowych trybach: *Overview*, *Data Laboratory* i *Preview* (il. 1). W oknie *Overview* wyświetla się graf z bieżącymi zmianami. Prawym przyciskiem myszy można go przesuwać w oknie, a rolką – skalować. W dolnej części okna są opcje wyświetlania węzłów (nazw) i krawędzi. W środku okna, na pionowej belce, są dostępne rozmaite narzędzia do modyfikacji atrybutów wybranych węzłów i krawędzi, np. koloru węzłów i otoczenia, rozmiaru. Można nawet wstawić lub usunąć pojedyncze obiekty. Po lewej stronie okna na dole znajduje się panel *Layout* umożliwiający wybór optymalnej konfiguracji sieci poprzez zastosowanie odpowiedniego algorytmu mapowania.

1 Odesłanie do strony internetowej przedstawia wersję aktualną w dn. 14.07.2017 r.

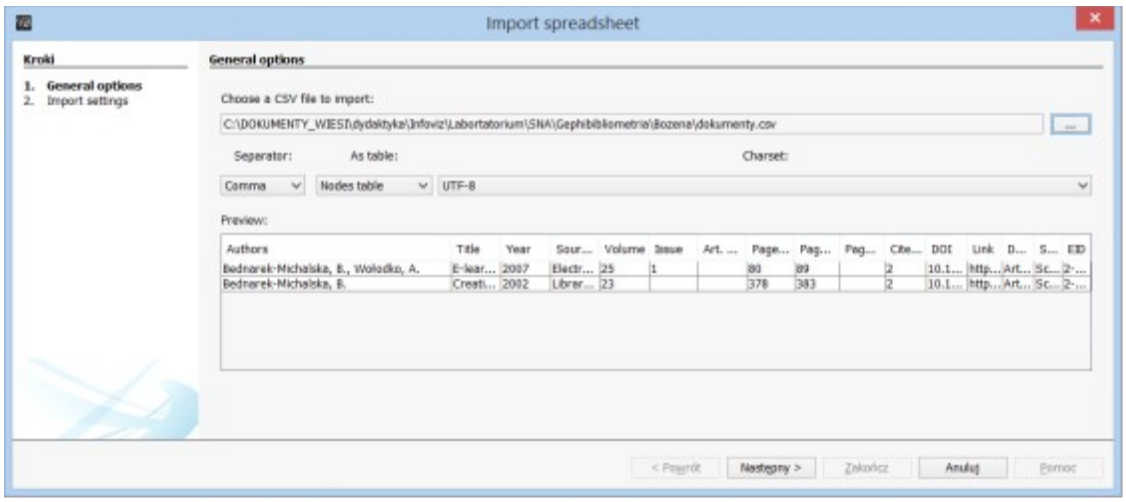


Il. 1. Obszar roboczy aplikacji Gephi.
Źródło: oprac. własne.

Algorytmy te bazują na różnych kombinacjach sił symulujących siły oddziaływania pomiędzy hipotetycznymi cząsteczkami jak również na sile scalającej wizualizowany układ w całość, tzw. sile grawitacji dośrodkowej. Stąd tak mocno różniące się wyjściowe wizualizacje sieci. Wybór layoutu wiąże się również z doborem w tymże panelu opcji, które dają najlepszy efekt wizualny z perspektywy poszukiwań analitycznych.

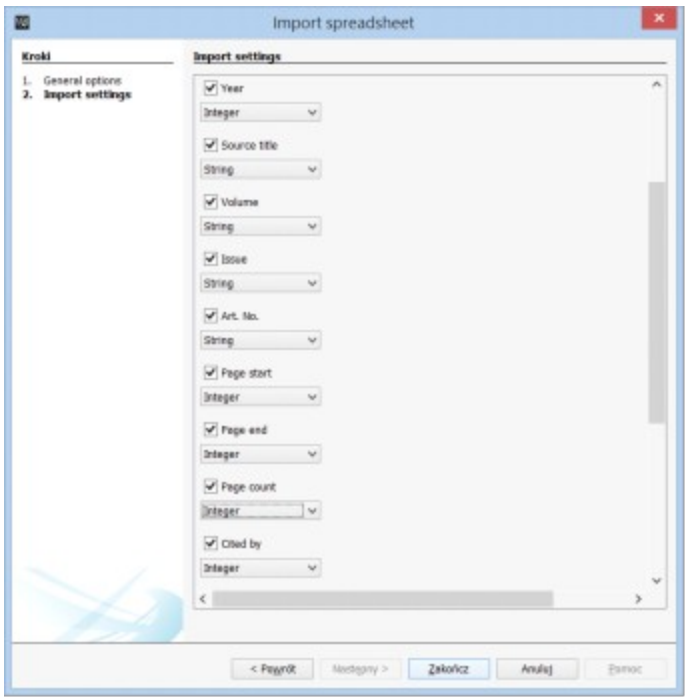
Panel *Appearance* (po lewej stronie ekranu) mieści funkcje do zmiany zbioru atrybutów węzłów i krawędzi na podstawie wskazanej zmiennej numerycznej lub nominalnej (tekstowej). Można w ten sposób zarządzać kolorem, rozmiarem, grubością obiektów i parametrami czcionki w nazwach. Nawet jeśli stworzymy idealny dla swoich celów graf, to końcowa wizualizacja, tuż przed eksportem do formatu graficznego, wymaga ponownego operowania atrybutami wizualnymi w panelu *Preview*. Panele *Filters* i *Statistics* (z prawej strony) zawierają mechanizmy obliczania podstawowych metryk w analizie SNA oraz wykorzystania powyższych w filtrowaniu obiektów sieci. Potrzebna jest tu jednak znajomości elementów analizy sieci społecznych oraz teorii sieci, a tak zaawansowany zakres treściowy nie mieści się w zakresie niniejszego tutorialu.

Natura naszych danych – dwa osobne pliki tekstowe – wymaga takiego ich przetworzenia w Gephi, żeby zostały one potraktowane jako zbiór charakteryzujący jedną sieć cytowanych dokumentów. Najwięcej wysiłku należy zatem włożyć w pracę z danymi w oknie *Data Laboratory*. Pierwszym krokiem jest otwarcie aktywnej zakładki *Nodes,* gdzie włączamy *Import Spreadsheet* i ładujemy plik „dokumenty.csv”. Istotne znaczenie mają: ustawienie właściwego separatora kolumn (w omawianym przypadku zastosowano przecinek (*Comma*)), selekcja listy na *Nodes table* oraz dobór odpowiedniego standardu kodowania znaków (dla Scopusa jest to UTF-8), co pokazano na il. 2.



II. 2. Aplikacja Gephi. Import pliku CSV z danymi o węzłach.
Źródło: oprac. własne.

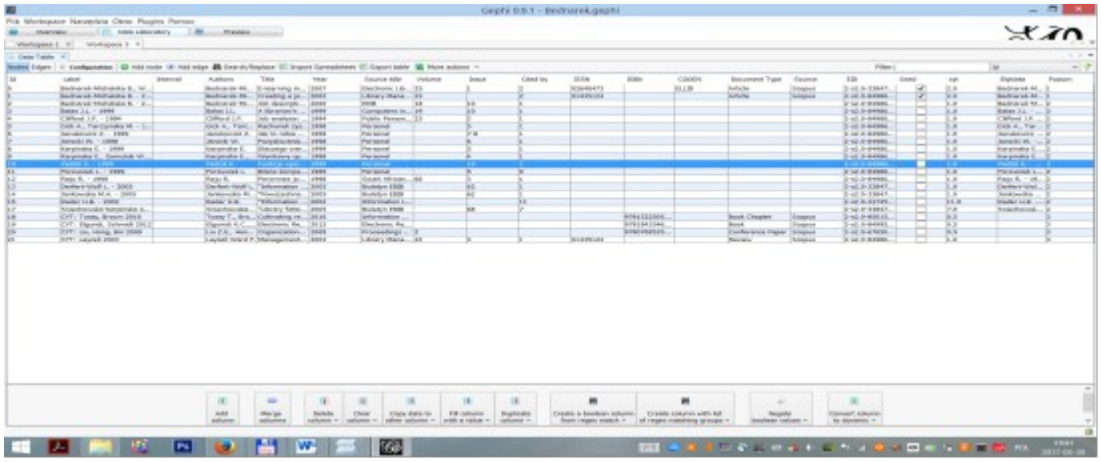
Następne kroki są bardzo istotne, jeśli chodzi o poprawne wczytywanie danych. Domyślnie program rozpoznaje wszystkie metadane jako dane typu tekstowego, a to komplikuje dalsze ich przetwarzanie. Pola numeryczne, które na pewno będą wykorzystane w analizach, należy zaznaczyć jako *Integer* (całkowite) lub *Double* (dla liczb dziesiętnych), tak jak pokazano na il. 3. Po operacji mamy dwa rekordy w tabeli i dwie kropki w oknie *Overview*. Brak etykiet (nazwisk autorów albo tytułów prac) przy węzłach spowodowany jest tym, że kolumna *Label* jest pusta. Można uzupełnić ten brak, wykorzystując dolny przycisk *Copy data to other column* i wybierając odpowiednią kolumnę jako źródło nazwy.



II. 3. Poprawione typy metadanych przy imporcie pliku CSV w aplikacji Gephi.
Źródło: oprac. własne.

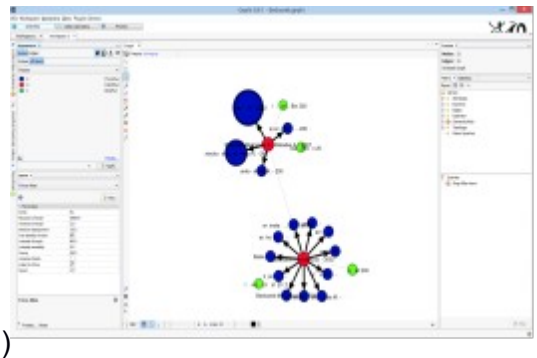
Następnie w ten sam sposób importujemy plik „cytowania.csv”, przy czym typy metadanych są automatycznie rozpoznawane. Ponownie można uzupełnić nazwy poprzez kopiowanie wartości kolumn. W wyniku w tabeli pojawiło się 6 rekordów z automatycznie nadanym identyfikatorem w pierwszej kolumnie.

Drugim krokiem jest zdefiniowanie powiązań pomiędzy dokumentami, czyli wyznaczenie krawędzi. Można to zrobić w Excelu, wypełniając tylko 3 kolumny następującymi danymi: identyfikator węzła źródłowego, identyfikator węzła docelowego oraz waga. Istnieje również możliwość wprowadzenia takich informacji bezpośrednio w arkuszu Gephi. Potrzebna jest do tego nowa kolumna. Może się jednak zdarzyć, że liczba kolumn osiągnęła maksimum, o czym informuje opcja *Display settings* (żółta żarówka z prawej strony nad tabelą), wówczas w tymże oknie można wyłączyć nieużywane kolumny. Tworzymy więc nową kolumnę (*Add column*) o typie *Boolean* (logiczny) i nazwie *Seed*. Odhaczamy w niej dwa dokumenty źródłowe, które są cytowane (il. 4). Ta kolumna będzie dla nas informacyjną w tworzeniu sieci połączeń. Kliknięcie prawym klawiszem myszy na rekordzie i wybór opcji *Link to nodes* spowoduje pojawienie się odpowiedniego rekordu w tabeli *Edges*. W towarzyszącym oknie dialogowym trzeba wybrać identyfikatory węzłów: źródłowego i docelowego oraz typ grafu: ukierunkowany lub nieukierunkowany. W przypadku operowania na wagach zazwyczaj wybiera się typ nieukierunkowany (*undirected*) – wówczas relacja pomiędzy węzłami jest symetryczna. Dla grafu ukierunkowanego krawędzie wyposażone są w strzałki, wskazujące kierunek przepływu informacji. Należy zauważyć, że strzałki są wyświetlane tylko dla prostych krawędzi, nie dla zakrzywionych. W danym przypadku posłużyliśmy się grafem ukierunkowanym ze względu na charakter danych.



Il. 4. Aplikacja Gephi. Tabela węzłów dokumentów trzech poziomów: dwóch źródłowych, cytowanych przez źródłowe oraz cytujących źródłowe.
Źródło: oprac. własne.

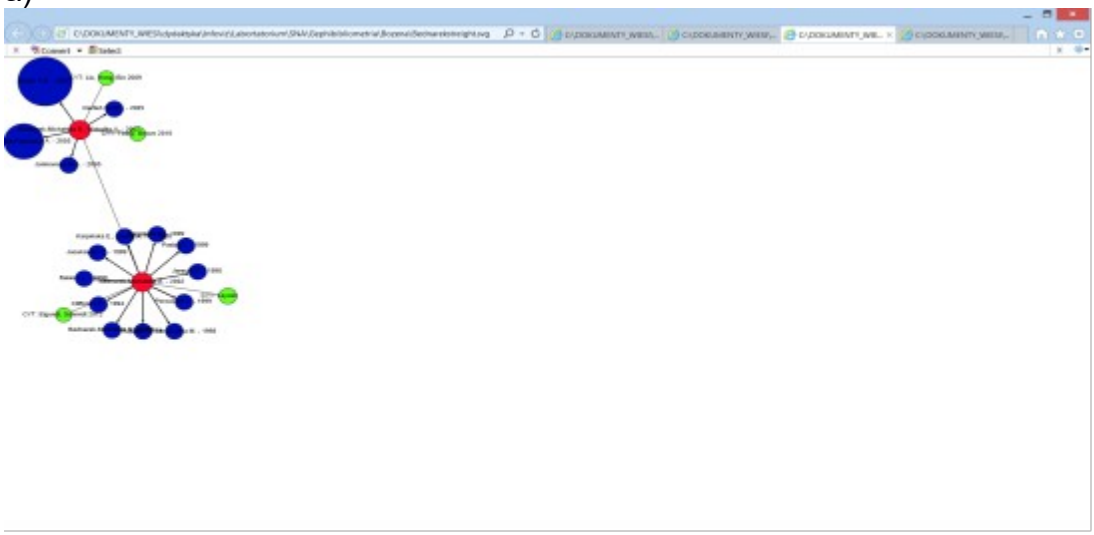
Naszą bazę dokumentów rozszerzyliśmy o artykuły cytowane. Zrealizowaliśmy to przez rozwinięcie każdego z dwóch dokumentów w bazie Scopus i pobranie w formacie CSV listy cytowanych dokumentów (*references*). Ostatecznie w tabeli ukazały się 22 rekordy. Dodatkowo stworzona kolumna *Poziom* informuje o poziomie dokumentu w strukturze cytowań: 1 – wskazuje na źródłowe dokumenty, 2 – na cytowania, 3 – na publikacje cytujące. Na podstawie wartości tej kolumny pokolorowano graf trzema kolorami. W oknie *Overview*, w zakładce *Appearance*, ustawiono parametry tak jak na il. 5a. Rozmiar węzłów sparametryzowano za pomocą zmiennej cytowania – *Cyt_by* (il. 5b).

a)  b) 

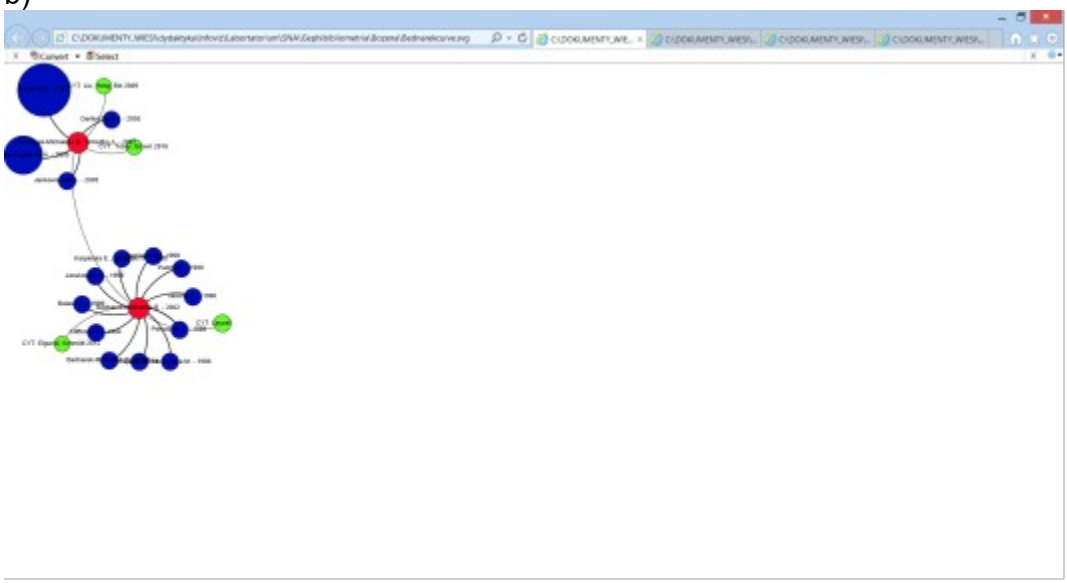
II. 5 Aplikacja Gephi. Ustawienie atrybutów koloru (a) oraz rozmiaru dla węzłów (b)..
Źródło: oprac. własne.

Po licznych próbach wybrano najbardziej czytelny layout *Yifan Hu Proportional*. Wynikowy graf jest przedstawiony na il. 6.

a)



b)



II. 6. Aplikacja Gephi. Wynik sieci cytowań wybranej osoby w postaci grafu ukierunkowanego z prostymi krawędziami (a) oraz zakrzywionymi (b). Źródło: oprac. własne.

Węzły dwóch podstawowych publikacji (poziom 1 w bazie), oznaczone na grafie kolorem czerwonym, skupiają wokół siebie cytowane publikacje (poziom 3 w bazie), tworząc symetryczne pierścienie (kolor niebieski). Węzły dokumentów, które cytują wymienione dwie publikacje (poziom 2) są koloru zielonego i ułożone są na zewnątrz względem cytowanych. Taki układ wyraźnie ukazuje ścieżki cytowań wybranego autora.

Podsumowanie

Tutorial miał na celu pokazanie bibliometrom, jak stworzyć sieć cytowań, posługując się danymi z bazy Scopus w środowisku Gephi. Gephi jest platformą służąca do analizy sieci społecznych. Polscy bibliometrzy, w oparciu o dane wieloskalowe, mogą skutecznie używać tego oprogramowania do wizualnych analiz, dzięki którym mogą pozyskać nową wiedzę o uczonych i ich wzajemnej współpracy. Przyciągającym elementem są tu także generowane, przy użyciu różnych algorytmów, atrakcyjne formy wizualizacji.

Tutorial zawiera informacje o standardach formatów plików sieciowych, ukazując wagę znajomości obsługi arkuszy kalkulacyjnych. Dane można przygotować w Excelu, którego z pewnością używają bibliometrzy, po czym zaimportować je do Gephi. Alternatywnym rozwiązaniem jest praca w arkuszu Gephi (*Data Laboratory*) od samego początku. Można w nim zbudować kompletną bazę cytowań lub współautorstwa. Oczywiście, jest to wygodny sposób pracy, gdy mamy do czynienia z liczbą kilkudziesięciu lub nawet kilkuset rekordów. Opanowanie dużych wolumenów – kilku- lub kilkudziesięciotysięcznych – wymaga zautomatyzowania procesu konwersji danych z baz Scopus lub WoS. Nad takim rozwiązaniem pracuje obecnie zespół naukowy, z udziałem autorki, w ramach projektu NCN.

Badania powstały w ramach projektu badawczego NCN 2014-2017 pt. Badanie struktury i dynamiki cyfrowych zasobów wiedzy za pomocą metod wizualizacji (ang. Information Visualization methods in digital knowledge structure and dynamics study) w Toruniu na Uniwersytecie Mikołaja Kopernika.

Bibliografia:

1. *BibExcel: a tool-box developed by Olle Persson* [online]. [Dostęp 8.07.2017]. Dostępny w: http://homepage.univie.ac.at/juan.gorraiz/bibexcel/.
2. *Gephi Features* [online]. The Gephi Consortium, 2017. [Dostęp 8.07.2017]. Dostępny w: https://gephi.org/features/.
3. *Gephi makes graphs handy* [online]. The Gephi Consortium, 2017. [Dostęp 8.07. 2017]. Dostępny w: https://gephi.org/.
4. OSIŃSKA, V. *WIZualizacja INFOrmacji: studium informatologiczne*. Toruń: Wydaw. Nauk. UMK, 2016. ISBN 978-83-231-3581-4.
5. OSIŃSKA, V., MALAK, P. *Dynamiczne sieci społeczne. Projektowanie i analiza.* W: SYSŁO, M.M., KWIATKOWSKA, A.B. (red.). *Wsparcie kształcenia informatycznego w szkołach: materiały pokonferencyjne*. Toruń: Wydaw. Nauk. UMK, 2017, s. 376-387. ISBN 978-83-231-3784-9.
6. *Pajek: analysis and visualization of large networks* [online]. [Dostęp 5.06.2017]. Dostępny w: http://mrvar.fdv.uni-lj.si/pajek/.