FOLIA 211

Annales Universitatis Paedagogicae Cracoviensis

Studia ad Bibliothecarum Scientiam Pertinentia XIV (2016)
ISSN 2081-1861
DOI 10.24917/20811861.14.4

Veslava Osińska, Piotr Malak, Bożena Bednarek-Michalska

Rozwój badań nad wizualizacją informacji

Wstęp do naukometrii wizualnej

Autorytet 90-letniego Eugene Garfielda, legendy w środowiskach naukometrycznych, rośnie w ostatnich latach. Jest to związane z poszukiwaniem uniwersalnych, wieloczynnikowych algorytmów parametryzacji dorobku naukowego. Analizując liczne fora i portale internetowe (np. Warsztat badacza[1]) oraz bieżące raporty ministerialne, można zauważyć, że na gruncie krajowym dochodzi do wielu błędnych interpretacji lub implementacji wskaźników wpływu, co z kolei prowadzi do negatywnych zjawisk, takich jak „punktoza"[2] ironicznie określana w Polsce, jako pogoń za punktami. Nasi zagraniczni koledzy nauczyli się przeciwdziałać takim praktykom, regularnie nagłaśniając ten temat w prasie akademickiej[3].

E. Garfield najczęściej kojarzony jest Institute for Scientific Information (dalej: ISI) i opracowaną przez tenże Instytut tzw. „listą filadelfijską". Przekształcenie ISI, w założeniu instytucji wyższej użyteczności, w komercyjne przedsięwzięcie, sygnowane przez Thomson Reuters, zmieniło podejście względem badań naukowych oraz dokumentowania ich wyników. W Wikipedii profesor Garfield jest określany jako biznesman[4], przypisuje się mu także wiele innych zasług, np.: wdrożenie pierwszych systemów informacyjno-wyszukiwawczych, wprowadzenie podstawowych dziś dla nauki indeksów: *Science Citation Index*, *Art and Humanities Citation Index*, założenie i redagowanie renomowanego czasopisma „The Scientist". Niewiele biogramów podaje, iż ten multidyscyplinarny uczony (chemik, bibliolog, lingwista) opracował i rozwinął system analityczny dla historii i rozwoju wybranego obszaru badań w oparciu o ścieżki cytowań.

W roku 1964, kiedy komputery nie były jeszcze powszechnie używane w środowisku naukowców, Garfield z kolegami sformułował pytanie badawcze, czy komputer

[1] E. Kulczycki, *Warsztat badacza – Emanuel Kulczycki*, [online] http://ekulczycki.pl/ [dostęp 30.06.2016].

[2] E. Kulczycki, *Post na blogu*, [online] https://www.facebook.com/emanuelkulczycki?fref=ts [dostęp 30.06.2016].

[3] *Leiden Manifest for Research Metrics*, [online] http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351 [dostęp 30.06.2016].

[4] *Wikipedia*. *Hasło Eugene Garfield*, [online] https://en.wikipedia.org/wiki/Eugene_Garfield [dostęp 30.06.2016].

może „opisać" historię nauki? Skoncentrowali się oni na naukoznawczym tle odkrycia struktury DNA, zainicjowanym w publikacji Jamesa Watsona i Francisa Cricka w 1953 roku. Z książki Isaaca Asimova *The Genetic Code* wyodrębnili 40 kluczowych prac, które doprowadziły do tego odkrycia. Ręcznie sporządzony wykres historiograficzny pokazywał równolegle powiązania historyczne, opisywane przez Asimova oraz powiązania cytowań tych 40 publikacji, wykrytych na podstawie bazy *Science Citation Index*. Obydwa podejścia wykazały wysoką zbieżność wyników[5].

40 lat później E. Garfield rozwinął tę koncepcję, tworząc aplikację HistCite[6], narzędzie, które automatycznie generuje chronologiczne tablice i historiografię kolekcji artykułów na określony temat. Ma ono pomóc naukowcom i bibliotekarzom w identyfikacji najważniejszych publikacji, wpływu wymienionych autorów, a przede wszystkim nakreśleniu ciągłości historycznej ewolucji/implementacji idei lub obszaru badawczego.

E. Garfield w swoich analizach bibliometrycznych wielokrotnie stosował graficzne prezentacje rozwoju nauki lub wybranych dziedzin naukowych. Na początku lat 90. wprowadził pojęcie **naukografii** (*scientography*)[7] i nazwał takie wizualizacje **naukogramami**. Jednak w naukoznawstwie te nazwy się nie rozpowszechniły. Obecnie najczęściej na określenie wizualizacji obszarów badawczych wraz z ich strukturami społecznymi używa się terminu *mapowanie nauki* (*Science mapping*)[8]. Wśród metod przetwarzania informacji do postaci graficznej dominują techniki *data mining, text mining*, algorytmy analizy danych oraz grafiki statystycznej. Tworzeniem map nauki zajmują się informatolodzy, bibliolodzy, informatycy, naukoznawcy, specjaliści od prezentacji informacji, graficy. Ten multidyscyplinarny kierunek badawczy rozwija się bardzo dynamicznie, a w dyskusjach naukometrycznych można spotkać ostatnio takie intuicyjne określenie, jak: „naukometria wizualna" (*visual scientometrics*)[9].

Niniejszy artykuł ma na celu zaprezentowanie możliwości wykorzystania zainicjowanego przez E. Garfielda nowego informatologicznego nurtu badań. Przedstawiono w nim serię przykładów wizualizacji dla kolekcji metadanych, które pochodzą z opisów bibliograficznych dokumentów dostępnych w polskich bibliotekach cyfrowych. Powstały one w trakcie badań przeprowadzonych w ramach grantu finansowanego z funduszy NCN 2013/11/B/HS2/03048 i dotyczyły analizy struktury zasobów cyfrowej wiedzy w Polsce przy wykorzystaniu nowoczesnych technik mapowania (zarówno 2, jak i 3D) informacji. Jednym z wymiernych efektów takich wizualizacji jest stworzenie serii funkcjonalnych map obrazujących zmiany zachodzące w polskiej nauce.

---

[5] E. Garfield, I. H. Sher, R. J. Torpie, *The Use of Citation Data in Writing the History of Science*, Pennsylvania 1964, s. 24.

[6] E. Garfield, *Historiographic mapping of knowledge domains literature*, „Journal of Information Science" 2004, 30(2), s. 119–145; E. Garfield, A. L. Pudovkin, V. S. Istomin, *Why do we need algorithmic historiography?*, „Journal of the American Society for Information Science and Technology" 2003, 54(5), s. 400–412.

[7] E. Garfield, *Scientography: Mapping the tracks of science*, „Current Contents: Social & Behavioural Sciences" 1994, 7(45), s. 5–11; E. Garfield, *Essays/Papers on Mapping the World of Science*, [online] http://garfield.library.upenn.edu/ mapping/mapping.html [dostęp 30.06.2016].

[8] K. Börner. *The Atlas of Science,* USA 2010, s. 10–13.

[9] Materiały konferencyjne konferencji *International Society of Scientometrics and Informetrics* z ostatnich 5 lat.

Od danych do map informacji

Z technologicznej perspektywy relacyjna baza danych składa się z jednej lub więcej tabel, które przechowują informacje o prezentowanym zjawisku bądź obiekcie w rekordach (wierszach). W podejściu statystycznym każdy rekord jest obserwacją jakiegoś wycinka rzeczywistości. Wielkość bazy danych jest przede wszystkim szacowana na podstawie liczby rekordów. W obrębie każdego rekordu właściwości (cechy) podmiotu opisywane są przy pomocy pól (kolumn). Tabele są powszechnie wykorzystywane w nauce do przechowywania i prezentowania danych zarówno strukturalnych, jak i obserwowanych. Posortowanie danych w tabeli znacznie ułatwia analizy zmian wartości poszczególnych cech. Kiedy wartości są małe, nie więcej niż na przykład pięcio-, siedmiocyfrowe, ludzka percepcja pozwala na wychwycenie zmian w kierunku spadkowym lub wzrostowym[10]. Jednakże w przypadku rejestracji danych w postaci liczb wielocyfrowych, dziesiętnych lub ułamkowych, układ tabelaryczny będzie mało przydatny w jakichkolwiek analizach wartości danych oraz ich zmian (trendów). Jeśli dodatkowo zamierzamy zbadać zachowanie wielu zmiennych, prezentowanych w kilku kolumnach, to oko ludzkie nie jest w stanie wychwycić ani zróżnicowania, ani korelacji. Podobnie kognitywno-percepcyjnym obserwacjom umykają wielkie wolumeny danych zapisane w postaci tabelarycznej czy jakiejkolwiek innej z wykorzystaniem tekstu. W takich sytuacjach pomocą w analizach i porównaniach danych są graficzne prezentacje ich wartości – ich wizualizacje. Jednym z rozwiązań wizualizacyjnych jest tzw. mapa cieplna (*heat map*), gdzie wprowadza się skalę kolorystyczną do kodowania istniejących wartości od najmniejszej do największej (por. Tab. 1 oraz Rys. 1).

Mapę cieplną można wygenerować np. w sieciowych programach do wizualizacji, jak również w popularnej aplikacji biurowej – MS Excel, używając narzędzia „formatowanie warunkowe". Intuicyjnym rozwiązaniem w prezentowaniu obserwowanych zmiennych jest zobrazowanie danych z tabeli przy pomocy wykresów, najczęściej słupkowych i kołowych. Ograniczeniem takiej wizualizacji jest zawsze liczba zmiennych, które jednocześnie można przedstawić za pomocą wybranego typu wykresu. Nadmiarowość wizualizowanych zmiennych w stosunku do wymiarów kartki wydruku (2D) lub przestrzeni 3D na ekranie monitora stanowi podstawowy technologiczny problem procesu mapowania. W krótkiej, kilkunastoletniej historii wizualizacji informacji problem ten rozwiązywano na dwa sposoby: poprzez modyfikowanie topologii wyjściowej, albo przez redukcję liczby zmiennych. W pierwszym przypadku rozciąga się przestrzeń, stosując technikę „rybiego oka". Tak działają przeglądarki graficzne, np. drzewo hiperboliczne albo Walrus, autorstwa Caidy (*Center for Applied Internet Data Analysis*)[11]. Do redukcji zmiennych

[10] F. Stephen, *Now you see it. Simple Visualization techniques and Quantitative Analysis*, CA, USA 2009, R. 3; V. Osińska, G. Osiński, A. B. Kwiatkowska, *Visuallization in Learning: Perception, Aesthetics and Pragmatism*, [w:] *Maximizinq Coqnitive Learninq throuqh Knowledqe Visualization*, red. A. Ursyn, Hershey, PA 2015, R. 13; G. A. Miller *The magical numer seven, plus or minus two: some limits of our capacity for processing information*, „Psychological Review" 2001, t. 101, nr 2, s. 343–352.

[11] *Walrus – Gallery: Visualization & Navigation*, Center for Applied Internet Data Analysis, [online] https://www.caida.org/tools/visualization/walrus/gallery1/ [dostęp 30.06.2016]; *Hyperbolic browser*, WikiViz, [online] http://www.wikiviz.org/wiki/Hyperbolic_browser [dostęp 30.06.2016].

stosuje się metody statystyczne (np. analiza głównych składowych – PCA, *Principal Components Analysis*, skalowanie wielowymiarowe) albo sztuczne sieci neuronowe (np. mapy Kohonena – SOM, *Self Organizing Maps*)[12]. Metody statystycznej redukcji zmiennych dotyczą ilościowej reprezentacji danych, przedstawianej jako macierz. Macierz prezentująca dane zawiera zazwyczaj bardzo dużo komórek o wartościach negatywnych (zero lub Null), oznaczających brak danej cechy w rekordzie/obserwacji. Redukcja zmiennych (ang. *matrix decomposition*) jest przeprowadzana w celu zaoszczędzenia miejsca na przechowanie oraz czasu celowych analiz danych. Analiza głównych składowych (PCA, zwana też *Karhunen-Loeve Transformation*-KLT) jest stosowana do macierzy prostokątnych (ilość kolumn różna niż ilość wierszy), czyli najczęściej występującej formy macierzy[13].

Za przykład zastosowania wizualizacji danych strukturalnych może posłużyć analiza finansowania konkursów Opus w trzech panelach dziedzinowych (dla sześciu edycji). Te same dane zostały zaprezentowane w postaci tabelarycznej oraz wygenerowanej na ich podstawie mapy cieplnej.

Tab. 1. Kwoty finansowania poszczególnych paneli konkursu OPUS, w kolejnych edycjach

	Opus 3	Opus 4	Opus 5	Opus 6	Opus 7	Opus 8
Razem	209 884 739	211 652 273	246 112 662	200 451 032	214 589 149	280 060 983
ST	101 849 442	103 770 136	101 689 025	80 480 673	92 973 210	123 464 341
NZ	80 927 024	75 967 789	100 144 579	80 185 031	88 119 960	122 261 050
HS	27 108 273	31 914 348	44 279 058	39 785 328	33 495 979	34 335 592

Dane tekstowe także można zmapować do postaci graficznej. Popularna obecnie metoda chmury słów (*tags cloud*) bazuje na obliczeniu częstości występowania danego wyrazu w korpusie tekstu. Przy tworzeniu takiej chmury pomija się w analizach słowa nie mające żadnej wartości informacyjnej, a najczęściej występują w danym języku, np. zaimki, liczebniki, spójniki, itp. Profesjonalne podejście do wizualizacji tekstu polega na konstruowaniu wektorowej reprezentacji terminów (w kolumnach) – dokumentów (w wierszach). Taki wektorowy model tekstu zapewnia konwersję do reprezentacji liczbowej, gdzie liczby opisują wagę danej cechy wyrazu, najczęściej frekwencji.

W celu dalszych analiz, w tym wizualizacji, zbiór danych należy sprowadzić do reprezentacji tabelarycznej. Taka postać jest wymagana w dalszych statystycznych przekształceniach i wizualizacji. W niniejszym artykule został opisany proces tabelaryzacji, analizy i wizualizacji danych tekstowych z opisów bibliograficznych. Autorzy dokonali przetwarzania i analiz kolekcji metadanych polskich bibliotek cyfrowych dostępnych na platformie Federacja Bibliotek Cyfrowych (FBC).

[12] V. Osińska, *Wizualizacja i wyszukiwanie dokumentów*, Warszawa 2010, s. 80–123.

[13] R. E. Madsen, L. K. Hansen, O. Winther, *Singular value decomposition and principal component analysis*, Raport techniczny 2004, [online] http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4000/pdf/imm4000.pdf [dostęp 30.06.2016]; G. Strang, *Introduction to linear algebra*, 4th Edition 2009, [online] http://math.mit.edu/~gs/linearalgebra/ [dostęp 30.06.2016].

Rys. 1. Mapa cieplna, obrazująca finansowanie konkursów Opus w trzech panelach dziedzinowych (dla sześciu edycji)

Źródło: opracowanie własne, 1 czerwca 2016 r.

FBC oferuje ujednoliconą platformę sieciową do przeszukiwania zasobów polskich bibliotek i wybranych repozytoriów cyfrowych[14]. W opisach dokumentów stosuje standard Dublin Core, udostępniając 15 pól opisu w opcji wyszukiwania zaawansowanego na portalu: począwszy od tytułu, twórcy, tematu, opisu, wydawcy, aż po źródło, język, zakres i prawa. Analizowany zbiór liczył ponad 2 mln rekordów. Ze względu na stawiany cel badawczy odfiltrowane zostały zasoby o charakterze naukowym. Dla pola „Type" została ściśle określona lista wartości nawiązujących do opracowań naukowych. Wynikowa baza danych zawierała *N* = 73 661 rekordów. Dane zostały sprowadzone do postaci tabelarycznej o następujących polach:

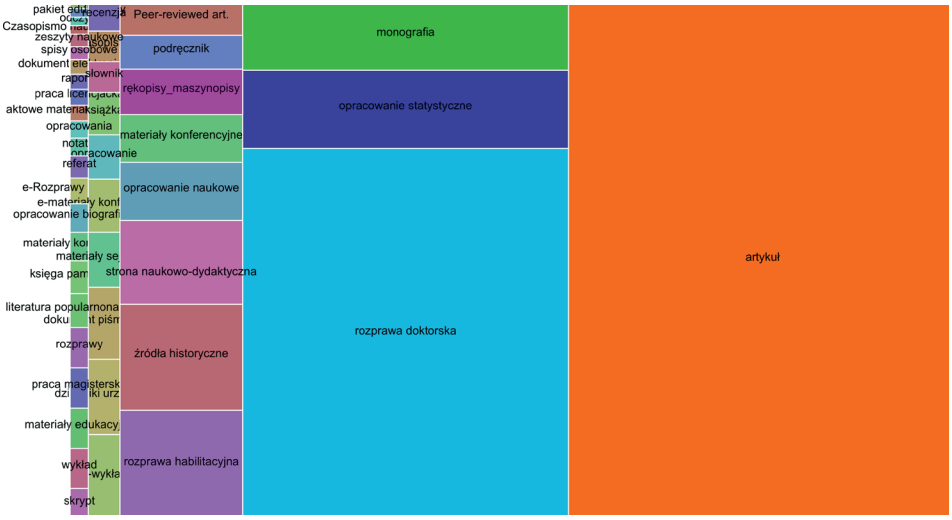
OAI Id; Tytuł; Autor; Współtwórca; Przedmiot; Język; Data publikacji; Data zdeponowania; Czas oczekiwania; Opis; Typ; Zasięg; Źródło; Prawa

W surowej postaci, czyli włączając różne formy syntaktyczne i językowe określeń typów dokumentów, lista typów liczyła ponad 150 wartości. Po grupowaniu semantycznym zbiór ten zredukowano dwukrotnie. Odrzucono rzadko występujące wpisy (o frekwencjach poniżej 20) i wynik zwizualizowano w postaci diagramu *treemap* (Rys. 2).

Oszacowanie pola powierzchni prostokątów pozwala na konkluzję, iż do najważniejszych typów w otrzymanej bazie danych zaliczyć można artykuły (31 200 obiektów), rozprawy doktorskie (18 229), opracowania statystyczne (3 934) i monografie (3 445).

Etapy czyszczenia, obróbki i wizualizacji metadanych z polskich bibliotek cyfrowych zostały opisane poniżej, a w kolejnym zamieszczone mapy wizualizacyjne prezentujące różne charakterystyki badanego zbioru danych.

[14] *Federacja Bibliotek Cyfrowych*, [online] http://fbc.pionier.net.pl/ [dostęp 30.06.2016].



Rys. 2. Lista wyselekcjonowanych typów dokumentów naukowych na diagramie treemap
Źródło: opracowanie własne.

Podstawowe problemy z danymi

Gromadzenie

Pierwszy i najbardziej czasochłonny etap wizualizacji danych skoncentrowany był wokół gromadzenia materiału badawczego. Dzisiaj, kiedy obowiązuje polityka upubliczniania danych i raportów dotyczących statystyk każdego obszaru życia społecznego[15], pobranie takich specjalistycznych informacji jest uproszczone. Polscy naukowcy mogą bez problemu skorzystać z globalnych agregatów danych, np. Web of Science, Scopus, Google Scholar i odfiltrować publikacje na określony temat. Z danych tych można wydzielić również naukowców badających konkretne zjawisko, albo też grono cytujących głównych twórców ustalonego w nauce kierunku, szkoły lub metodologii.

Użytkownicy bardziej zaawansowani w technologiach komputerowych wybierają często rozwiązania alternatywne do publicznych baz danych. Są nimi narzędzia dedykowane kolekcjonowaniu danych ze stron internetowych, tak zwane *Web scraper-y*. W naukometrii (a ściślej altmetrii/webometrii) wykorzystywane są statystyki odwiedzin serwisu naukowego, mierniki alternatywne, takie jak: ilość pobrań plików, przeglądy serii zasobów sieciowych, jak również zestawienia korespondencji mailowej pomiędzy osobami w badanej grupie, o ile nie są chronione prawami autorskimi.

Ogólnie rzecz ujmując, każdy język skryptowy (JavaScript, PHP) bądź programowania (Python, R, Java) dostarcza biblioteki pozwalające na skanowanie, pobieranie

[15] Takim sztandarowym przykładem jest stworzony przez fińskich nauczycieli portal: *GapMinde,* [online] https://www.gapminder.org/r [dostęp 30.06.2016]. Ta bardzo popularna aplikacja sieciowa w oparciu o globalne dane statystyczne z ostatnich 200 lat umożliwia analizy rozwoju społeczno-gospodarczego w skali dowolnego kraju i całego świata.

i co ważne czyszczenie danych surowych. Pobieranie danych i późniejsza ich organizacja wykorzystuje struktury tabelaryczne. Kolumny prezentują wyodrębnione w kodzie HTML bądź XML pola, wiersze każdą jednostkową informację o podmiocie. Format XML znacznie ułatwia prace nad gromadzeniem danych, ponieważ w samym swym założeniu przechowuje strukturę (drzewiastą) ich opisu. W przypadku języków znacznikowych, jak np. XML czy HTML, konieczne jest m.in. oczyszczenie surowych danych ze znaczników, o czym będzie mowa dalej.

Czyszczenie

Kolejnym etapem prac jest oczyszczenie pobranych danych i znormalizowanie ich reprezentacji. W przypadku metadanych z bibliotek cyfrowych mamy do czynienia z kilkoma problemami natury technologicznej. Dla przykładu jako znak oddzielający poszczególne pola metadanych część bibliotek stosuje cudzysłów – '"', zaś część pipeline – '|'. Różne bywają także znaki końca wiersza EOL (*EndOfLine*), w większości przypadków są to tradycyjne już CRLF (*carriage return*, *line feed*) domyślnie stosowane przez aplikacje systemu Ms Windows, trafiały się jednakże pliki z użytym znakiem LF, z systemów unixowych. Kolejne różnice, w zapisie plików z danymi, dotyczyły kodowania znaków. Większość danych z bibliotek cyfrowych jest obecnie kodowana w standardzie UTF-8, ale trafiają się również zapisane w standardzie Windows-1250. To wywołuje konieczność konwersji wszystkich plików z danymi do wspólnego kodowania oraz normalizacji do jednolitego prezentowania znaków kontrolnych, takich jak np. znak końca wiersza czy znak oddzielający zawartość poszczególnych pól. Innym krokiem normalizacyjnym jest decyzja o ujednoliceniu zapisu wielkich liter – na potrzeby indeksowania i wyszukiwania informacji stosuje się konwersję dużych liter na małe.

Dla języków fleksyjnych, a takim jest język polski, przeprowadza się również ujednolicenie zapisu postaci graficznej wyrazu. Najlepszym przybliżeniem, zachowującym znaczenie wyrazu, jest lematyzacja, czyli zastąpienie wszystkich wyrazów w zbiorze danych odpowiadającymi im formami podstawowymi (lematami). Są to np. formy mianownika liczby pojedynczej dla rzeczowników, czy pierwsza osoba czasu teraźniejszego dla czasowników. W omawianych badaniach w celu lematyzacji analizowanych treści skorzystano z zasobów dostępnych w ramach infrastruktury CLARIN-PL[16]. Na podstawie funkcji częstości lematów można wyznaczać wiele zbiorów cech reprezentujących dokumenty w zbiorze (tu: opis bibliograficzny).

Po znormalizowaniu sposobu zapisu danych należy te dane zaprezentować w formie tabeli. Konieczna jest tutaj analiza struktury danych dostępnych zasobów, w celu ustalenia wspólnych cech, występujących w każdym podzbiorze. W przypadku omawianych badań cechami były pola standardu DublinCore opisu bibliograficznego dokumentu, zaś podzbiorami były zestawy metadanych z poszczególnych bibliotek cyfrowych.

[16] *Clarin PL – Polska część infrastruktury naukowej CLARIN ERIC*, [online] http://ws.clarin-pl.eu/demo2/tager.shtml [dostęp 30.06.2016]. Tager jest programem który opisuje poszczególne elementy tekstu metainformacjami, czyli informacjami o funkcji i klasie gramatycznej wyrazu.

Przetwarzanie

Oczyszczone i znormalizowane dane są gotowe do dalszych analiz, które zaczyna się od trywialnych operacji sortowania i filtrowania w celu wyodrębnienia z całego zbioru rekordów spełniających zadane kryteria. Następnie porównuje się uzyskane dane, wyszukuje tendencje i relacje zachodzące pomiędzy nimi. Ostatnim etapem, szczególnie istotnym w przypadku wielkoskalowych zbiorów danych, jest wizualizacja wyników, obrazująca odkryte zależności i związki pomiędzy danymi. Dla zdecydowanej większości operacji przetwarzania i porównywania danych konieczne jest zaprezentowanie ich w postaci liczbowej. W tym celu, co zostało już opisane, wybiera się odpowiednią reprezentację danych. Zazwyczaj jest to jakaś funkcja częstości danych w zbiorze, np. frekwencja występowania lematów, częstość występowania danej części mowy lub klasy gramatycznej wyrażeń itp. Przykładowo, przy analizie współautorstwa obiektami są artykuły, a cechą jest para współautorów lub cytowania danego artykułu (w tym przypadku można dodatkowo uwzględnić współcytowania, jako cechę rozszerzoną). W mapowaniu współcytowań cechami są pozycje bibliograficzne, a zlicza się dokumenty cytujące parę artykułów (DCA – *Document Cocitation Analysis*) lub parę autorów (ACA – *Author Cocitation Analysis*). Ten etap wymaga dużej mocy obliczeniowej uzależnionej od rozmiaru bazy danych.

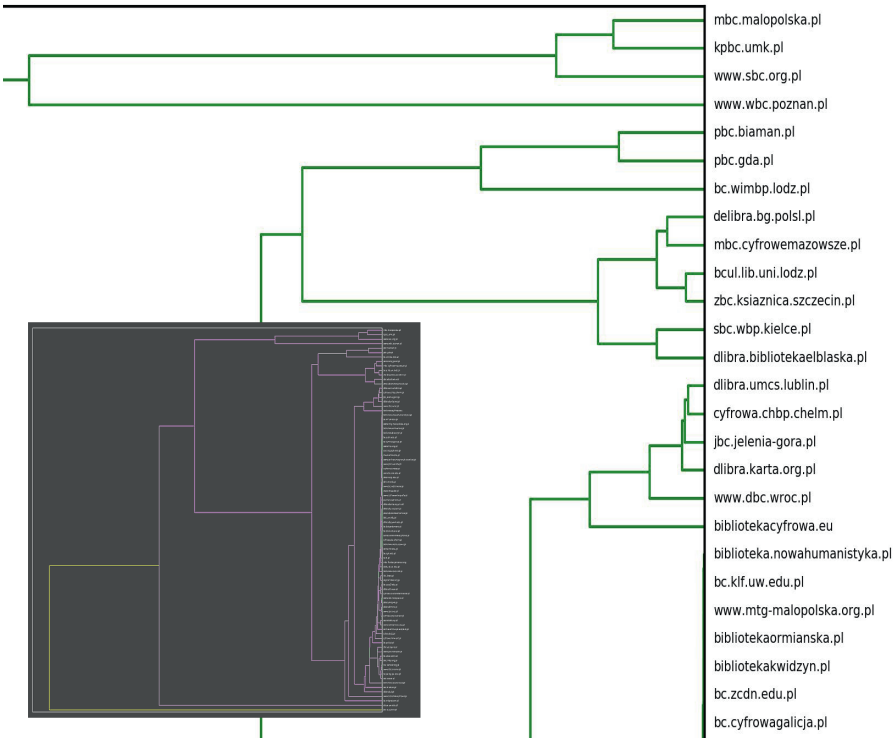
W przypadku badania intensywności współpracy naukowej można posłużyć się analizą współautorstwa. Na przykładzie metadanych z polskich bibliotek cyfrowych proces rozpoczynał się od wyodrębnienia dokumentów z podanym autorstwem. Na potrzeby analizy współpracy naukowej konieczne było filtrowanie uzyskanego zbioru w celu wskazania prac autorstwa pracowników nauki polskiej. Następnie zostały oszacowane liczebności podzbiorów ze względu na liczbę współautorów, i ostatecznie można było wizualizować uzyskane dane. Nazwy autorów do filtrowania weryfikowane były z listą Ośrodka Przetwarzania Informacji (OPI) „Ludzie Nauki". Wizualizacja intensywności współpracy naukowej w Polsce przedstawiona jest na rysunkach 5 oraz 6 w dalszej części niniejszego artykułu.

Grupowanie/kategoryzowanie

Dane zaprezentowane w postaci przeliczalnej, jako wartości liczbowe poszczególnych cech, można również automatycznie grupować. Zestaw liczb opisujących cechy jednego obiektu (jeden rekord w bazie danych) tworzy wektor. Porównując wektory można wskazać poziom ich podobieństwa (podobna wartość, kierunek czy zwrot wektora w przestrzeni). W przypadku metadanych z polskich bibliotek cyfrowych dokonano m.in. analizy podobieństwa bibliotek ze względu na typy (oraz ich liczebność) dokumentów w zbiorach. Wyniki takiego grupowania, w postaci dendrogramu, prezentuje rysunek 3.

Kategorie wizualizacji informacji. Studium przypadku

Typ analizowanych danych wpływa na wybór metody wizualizacji. Metody te stanowią pokaźną kolekcję i mają długą historię. W literaturze fachowej można spotkać opracowania, prezentujące znane techniki wizualizacji informacji na



Rys. 3. Klastry podobieństwa bibliotek cyfrowych na podstawie typów dokumentów
Źródło: opracowanie własne, 1 czerwca 2016 r.

przestrzeni ostatnich 200 lat[17]. Na gruncie polskim Veslava Osińska podjęła się opracowania logicznej systematyki współczesnych metod wizualizacji, co opisała w swoich publikacjach[18]. Proponuje, żeby uporządkowanie mocno zróżnicowanych metod oprzeć na typach analizowanych danych i/lub rodzaju informacji (np. treść, idee). Podążając tym wątkiem, w niniejszym artykule wyodrębniono podstawowe kategorie metod wizualizacji, wspierając je konkretnymi przykładami charakteryzującymi właściwości kolekcji FBC.

Wykresy

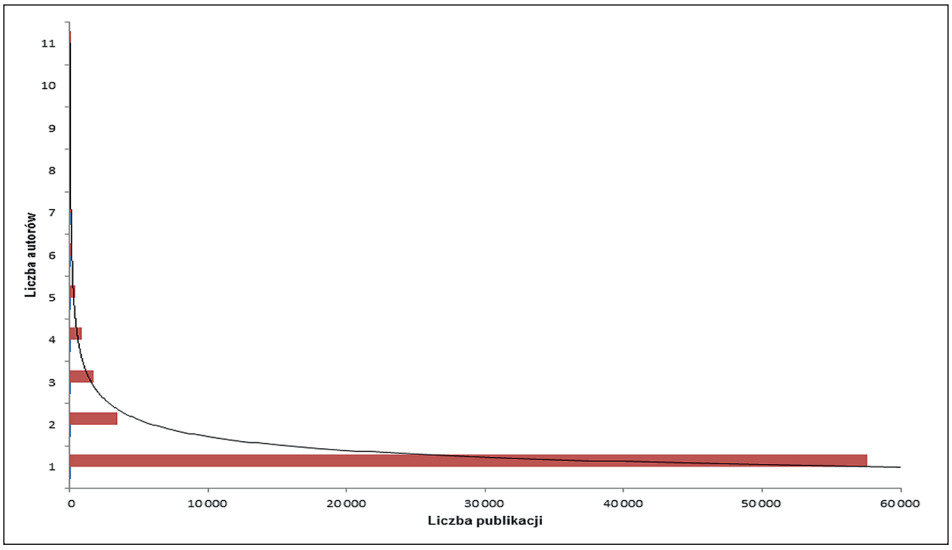
W prezentowanym zbiorze metadanych z polskich bibliotek cyfrowych tylko trzy pola zawierają liczbowy typ danych, reszta to wartości tekstowe. Jedną z możliwości zwizualizowania w zwarty sposób licznych kategorii nominalnych (tu typów dokumentów naukowych) jest diagram *treemap*, prezentuje rysunek 2.

Z bazy danych, składającej się przeważnie ze zmiennych nominalnych, można uzyskać zestawienia liczbowe dzięki technikom statystycznym. Najprostszym

[17] R. Lengler, M. Eppler, *Visualisation methods Periodic Table*, 2007, [online] http://www.visual-literacy.org/periodic_table/periodic_table.html [dostęp 30.07.2016]; J. Schabish, *The Graphic Continuum*, [online] http://visual.ly/graphic-continuum [dostęp 30.06.2016].

[18] V. Osińska, *Wizualizacja informacji. Studium informatologiczne*, Toruń 2016, s. 87–120.

sposobem jest zaprezentowanie danych nominalnych jako funkcję ich frekwencji. Kolejnym przykładem takiego przekształcenia informacji z metadanych polskich bibliotek cyfrowych może być analiza współautorstwa dokumentów. W bazach danych edytowanych ręcznie często zdarza się zaszumienie właściwych danych z powodu błędów operatora. Takimi są z pewnością wpisane inicjały autorów lub niezrozumiałe skróty. Ostatecznie po odrzuceniu rekordów z błędami lub brakiem wartości w polu *Creator* analizie statystycznej poddano 79% pierwotnego zbioru. Rozkład autorstwa indywidualnego i zespołowego dokumentów dostępnych na platformie FBC, prezentuje wykres słupkowy na rysunku 4. Linia interpolacyjna potwierdza idealny, potęgowy rozkład dla zespołowego współautorstwa naukowców: $f(x) \sim x^{-\alpha}$.



Rys. 4. Współautorstwo publikacji naukowych. Krzywa trendu wskazuje na rozkład potęgowy
Źródło: opracowanie własne, 1 czerwca 2016 r.

Trzy wspomniane typy numeryczne w metadanych odnoszą się do pól: daty publikacji dokumentu, daty zdeponowania go w bibliotece cyfrowej oraz różnicy pomiędzy tymi datami, którą autorzy nazwali czasem uwolnienia zasobów (CUZ). Do analizy zależności pomiędzy datą publikacji, a datą uwolnienia zasobu najlepiej wykorzystać histogram – wykres słupkowy pokazujący rozkład częstości badanych wartości.

Według autorskiej koncepcji, rozkład CUZ w czasie może wskazywać na rodzaj biblioteki ze względu na kryterium organizacyjne[19]. Masowa digitalizacja zbiorów jest ograniczona przede wszystkim prawami autorskimi, które pozwalają na upublicznienie obiektu po upływie 70 lat od śmierci autora lub od czasu powstania dzieła. To oznacza, że w bibliotekach cyfrowych powinny przeważać dokumenty pochodzące sprzed co najmniej 70 lat. Wykres zaprezentowany na rysunku 5 dowodzi, że faktycznie w kolekcjach cyfrowych istnieje trend odwrotny od intuicyjnego. Nie ma przewagi kolekcji z ponad 70–letnimi dokumentami, natomiast całość zdominowana

[19] V. Osińska, B. Bednarek-Michalska, P. Malak, *Charakterystyki czasowe deponowania zasobów w nakreśleniu dynamiki rozwoju i profilu polskich bibliotek cyfrowych*, „Zagadnienia naukoznawstwa" 2016 (w recenzji).

jest przez źródła współczesne (wartości zerowe na osi *X*). Taki stan rzeczy dowodzi, że współcześni naukowcy coraz chętniej deponują swoje publikacje z ostatnich lat, co ma odwzorowanie w kolekcjach cyfrowych bibliotek.

Mapy

Tradycyjnie mapa kojarzona jest z kartograficznym odwzorowaniem obszaru geograficznego na płaszczyźnie. W procedurze mapowania informacji skupiamy się na wykorzystaniu nie współrzędnych geograficznych, lecz danych abstrakcyjnych, mających źródło w Internecie, rozległych bazach danych lub innych ich źródłach, których podstawową cechą jest masowość. Mogą to być statystyki odwiedzin danej witryny czy aktywności użytkowników specjalistycznego serwisu albo dane ruchu internetowego itp. Wynikowa przestrzeń może być zarówno płaska, jak i w 3D. Kluczowym zagadnieniem w tym procesie jest odwzorowanie cech obiektów i relacji pomiędzy nimi na wyznaczoną topologię przestrzenną. Może ona łączyć rzeczywistą geografię z korelacjami opisanymi ilościowo za pomocą atrybutu graficznego, tak jak przedstawia to rysunek 6a.

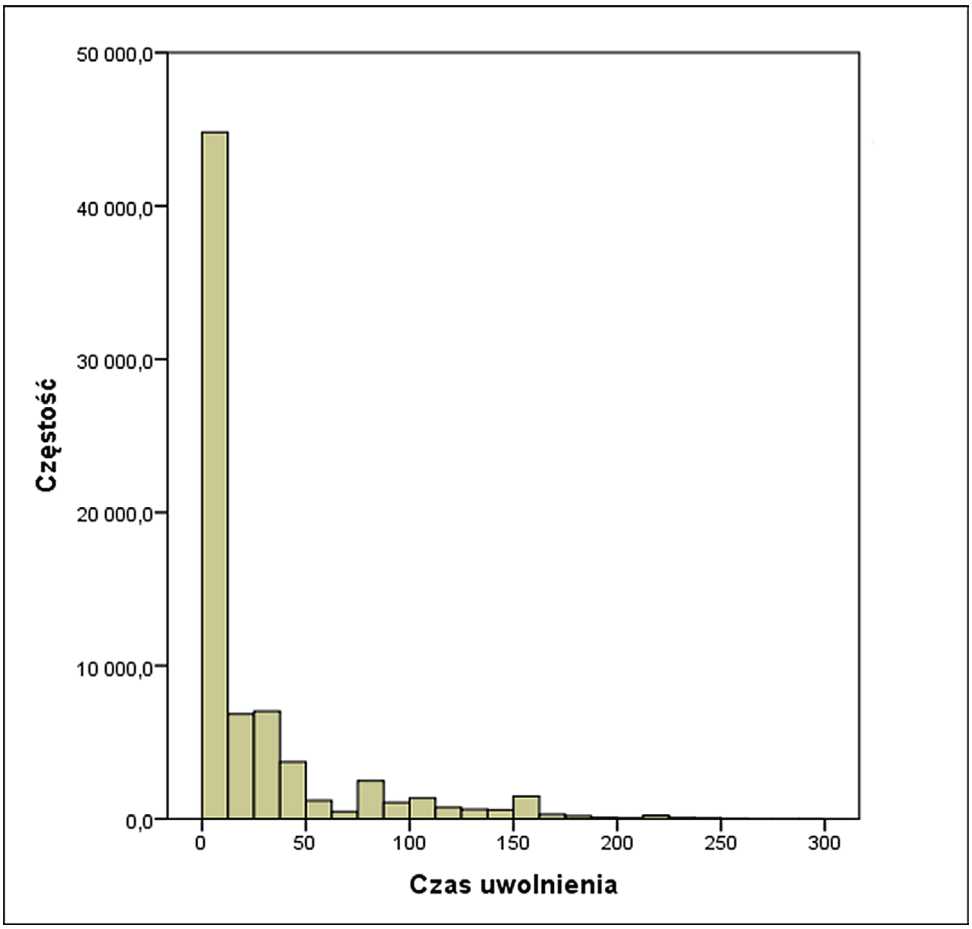
Rekordy, w których zidentyfikowano autora, poddano filtrowaniu za pomocą innej bazy, zawierającej informację o afiliacji pracowników naukowych, ponieważ nie wszyscy publikujący figurują w tzw. bazie OPI. Na tej liście znajdują się osoby ściśle związane z awansem naukowym i są poddawane systematycznej parametryzacji. Na rysunku 6a pokazano współpracę maksymalnie dwóch autorów przy publikowaniu artykułów naukowych. Grubość łączeń pomiędzy miastami służy za atrybut graficzny częstości współautorstwa (odrzucono wartości 1 i 2 – jedna i dwie wspólne prace), zaś wielkość kółka jest proporcjonalna do publikacyjnej aktywności naukowców z danego miasta. Podobnie na diagramie bąbelkowym (rys 6b), pole powierzchni kół wskazuje na potencjał miasta w zakresie współpracy większych zespołów akademickich, składających się z więcej niż 2 badaczy.

Grafy

Tak jak zaakcentowano powyżej, wizualizacja, a w szczególności jej współczesne, wyrafinowane percepcyjnie metody, pozwalają na wykrycie korelacji pomiędzy badanymi obiektami, znalezienie wspólnych cech, które wcześniej były ukryte, a które ostatecznie dostarczają nowej wiedzy o obserwowanych zjawiskach. W nauce może to prowadzić do odkrycia paradygmatów naukowych[20].

Sprawdzonym, efektywnym sposobem prezentowania związków pomiędzy licznymi danymi jest wykorzystanie grafów. Do stworzenia takich struktur, składających się z wielu wierzchołków i wzajemnych połączeń, potrzebna jest informacja o liczbie tych wierzchołków oraz sile wiązania, czyli wadze. Takie wizualizacje za pomocą grafów (tzw. sieciowe) są obecnie wykorzystywane do przedstawienia

[20] Ch. Chen, J. Kuljis, *The rising landscape: a visual exploration of superstring revolutions in physics*, „Journal of the American Society for Information Science and Technology" 2003, 54(5), s. 435–446; V. Osińska, *Wizualizacja paradygmatów naukowych*, „Zagadnienia naukoznawstwa" 2012, nr 48, s. 205–220.



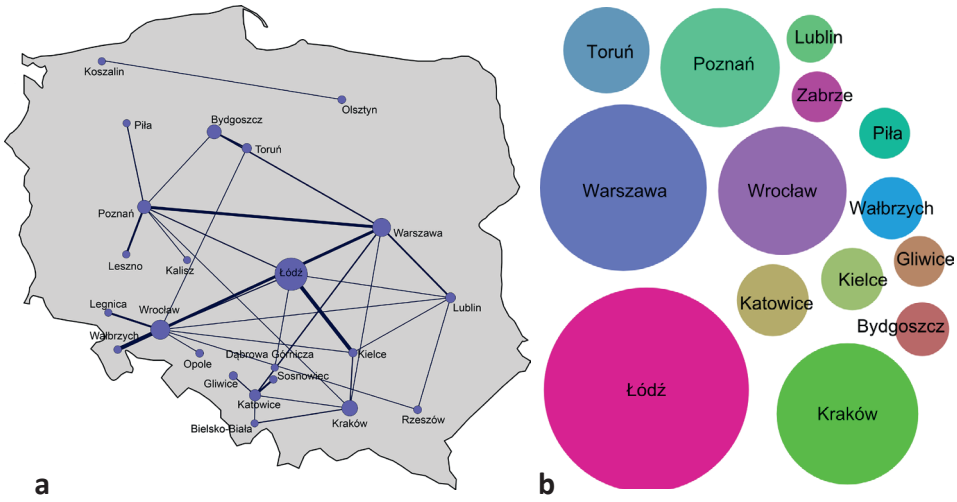
Rys. 5. Histogram czasu uwolnienia zasobów o charakterze naukowym na podstawie danych Federacji Bibliotek Cyfrowych

Źródło: opracowanie własne, 1 czerwca 2016 r.

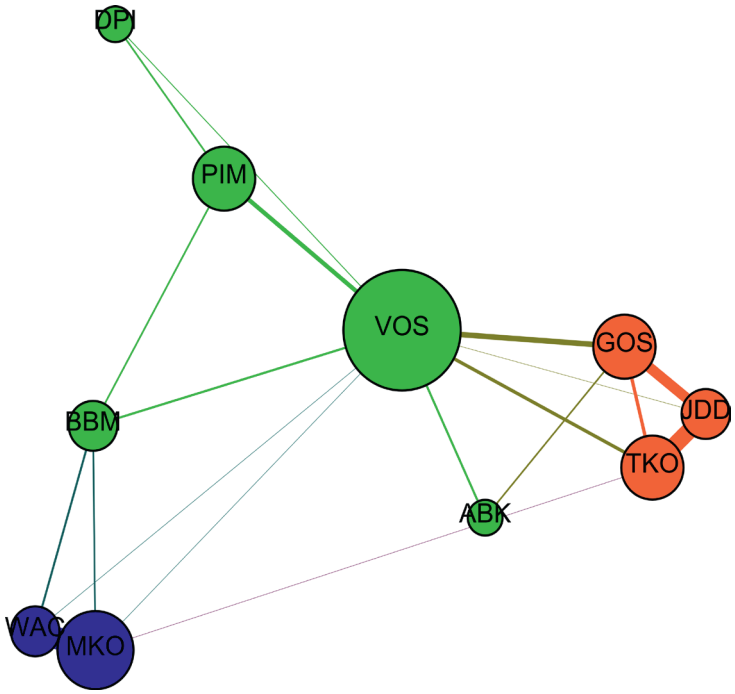
skomplikowanych struktur dynamicznych, podobieństw badanych obiektów oraz współpracy ludzi w obrębie społeczności dowolnego formatu: zespołu, organizacji, kraju, federacji, świata. Wynikowe, kolorowe, nawiązujące do fraktali prezentacje są niezwykle estetyczne, co tłumaczy ich popularność w każdej dziedzinie wiedzy.

Sieć współpracy wybranych członków nieformalnej grupy toruńskich naukowców wykorzystujących metody wizualizacji danych w swoich badaniach VIS-team[21] przedstawiony jest na rysunku 7. W wizualizacji wykorzystano dane ze społecznościowego portalu dla naukowców ResearchGate. Liczba wspólnych publikacji określa wagę połączeń pomiędzy autorami: im grubsza linia, tym mocniejsza, co jest również odwzorowane w odległości pomiędzy wierzchołkami. Wielkość kółka zależy od liczby kompetencji (*skills*), zarejestrowanych na profilu przez autora i potwierdzonych przez innych użytkowników portalu, co w założeniu daje obiektywny miernik

[21] Więcej o grupie i badaniach można poczytać na stronie: Wizualizacja informacji, [online] http://www.wizualizacjainformacji.pl/onas.php [dostęp 30.07.2016].



Rys. 6. Geograficzny rozkład współautorstwa w parach (**a**) oraz diagram bąbelkowy ilustrujący miasta autorów publikujących zespołowo w większych grupach (**b**)
Źródło: opracowanie własne, 1 czerwca 2016 r.



Rys. 7. Wizualizacja współpracy wybranych członków grupy VIS-team na podstawie danych o współautorstwie i umiejętności na portalu ResearchGate. Każdej specjalizacji badaczy odpowiada odcień: dwóm bibliologom z lewej strony – najciemniejszy, pięciu informatologom w środkowej części – najjaśniejszy, pozostali trzej są kognitywistami
Źródło: opracowanie własne, 1 czerwca 2016 r.

profesjonalizmu w danym zakresie. Na wagę dodatkowo wpływa liczba wspólnych kompetencji. Ta sama lub podobna tematyka badań zacieśnia współpracę. Kolorem (odcieniem) zakodowane są specjalizacje badaczy w tym multinterdyscyplinym zespole: bibliolodzy (ciemny, z lewej), informatolodzy (najjaśniejszy), kognitywiści (po prawej). Otrzymana wizualna konfiguracja pozwala na ocenę potencjału współpracy wewnątrz grupy, możliwe ścieżki poszerzenia lub zmiany specjalizacji, jak również prognozowanie przyszłych zmian. Z pewnością wizualizacja większej liczby obiektów dostarczyłaby bardziej wartościowych obserwacji, jednak to zadanie nie jest łatwe ze względu na konieczność zautomatyzowania procesu zbierania danych, których udostępnienie nie leży w interesie komercyjnego portalu.

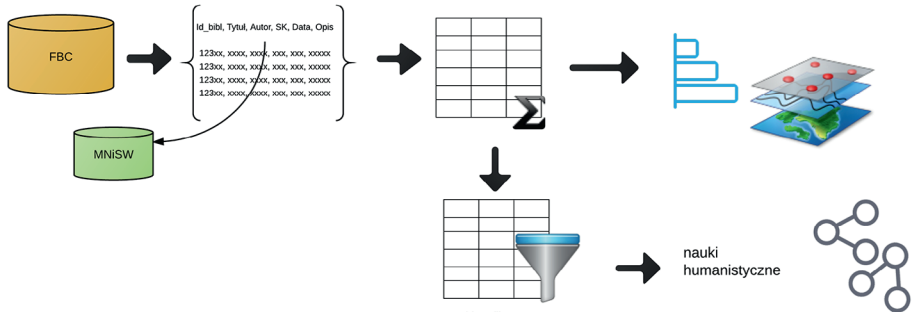
Diagramy i infografiki

Diagramy jako formy komunikowania koncepcji i idei stanowią stały element literatury naukowej. Pierwsze udokumentowane grafiki, mające kształty zaczerpnięte z obserwacji przyrody (np. drzewo, gałęzie, okręgi), można bez większego trudu znaleźć w historycznych starodrukach[22]. Dzisiejsze diagramy i schematy są z łatwością projektowane w edytorach tekstowych i graficznych. Z reguły wykorzystują one gotowe szablony, których podstawą jest elementarna geometria. W celu pokazania proporcji składowych całości stosowane są z podziałem na segmenty okręgi, trójkąty, kwadraty. Aby zademonstrować pokrywające się wątki lub tematy, niezmiennie stosuje się diagram Venna, ilustrujący logiczne zależności pomiędzy zbiorami. Hierarchię informacji najlepiej ująć za pomocą drzewa (dendrogramu). Harmonogram planowanych zadań tworzy się za pomocą diagramu Gantta. W aplikacjach biurowych automatycznie można stworzyć schemat koncepcji lub wiedzy o procesach i właściwościach analizowanego fragmentu rzeczywistości.

Obecnie, kiedy mocno wzrosło znaczenie estetyki przekazu oraz dzięki powszechnej dostępności programów graficznych, jak również alfabetyzacji uczniów w tym zakresie (edytor Gimp i Inkscape jest włączony do programu nauczania informatyki w szkołach ponadgimnazjalnych), „euklidesowe" diagramy ulegają modernizacji stylistycznej i przypominają infografikę. Sztywne formy urozmaicane są ikonkami, klipartami, a nawet szkicami wykonanymi odręcznie. Te ostatnie, ukierunkowane na ludzką inwencję i szybkość interakcji, definiują nowy trend przekazywania wiedzy w biznesie, tzw. „myślenie wizualne"[23]. Odpowiednim przykładem jest infografika ilustrująca schemat bieżącego eksperymentu na rysunku 8. Pokazuje on kolejność etapów, środki i cele badawcze. Robocza baza danych tworzona jest z pierwotnej (FBC), zredukowanej poprzez filtr zasobów ludzkich (MNiSW) oraz poddawana jest obróbce statystycznej, żeby ostatecznie uzyskać charakterystyki wizualne różnych aspektów badanych zasobów cyfrowej wiedzy. Jednym z celów jest poklasyfikowanie analizowanych metadanych według dziedziny naukowej i tym samym wyselekcjonowanie obiektów odnoszących się do humanistyki i nauk społecznych.

[22] M. Lima, *The Book of Trees. Visualizing Branches of Knowledge*, New York, 2014; S. Weingart, *Diagrams of knowledge*, Blog 2013, [online] http://www.scottbot.net/HIAL/?page_id=39166 [dostęp 30.06.2016].

[23] K. Jóźwik, Sz. Zwoliński, *Myślenie wizualne w biznesie. Ty też potrafisz rysować*, Warszawa, 2015; V. Osińska, G. Osiński, A. B. Kwiatkowska, *Visuallization*…



Rys. 8. Schemat eksperymentu uzyskiwania, przetwarzania, filtrowania i wizualizacji metadanych bibliotek cyfrowych

Źródło: opracowanie własne., 1 czerwca 2016 r.

Infografikę zaczęto wykorzystywać także jako abstrakty w literaturze naukowej, tak jak robi to od lat wydawnictwo Elsevier[24]. Graficzne streszczenie może służyć jako podsumowanie najważniejszych ustaleń i uzyskanych wyników artykułu naukowego. Redaktorzy czasopisma „Informetrics" zachęcają autorów do załączenia abstraktu graficznego, który umożliwiłaby szybkie zrozumienie głównego wątku publikacji. Można również zaobserwować, że studenci chętniej przedstawiają wyniki własnych badań w formie infograficznej zamiast prezentacji PowerPoint[25]. Wydaje się, że jest to bardziej efektywny sposób komunikowania wiedzy w środowisku młodych osób, aktywnie korzystających z nowych mediów, niż linearny, (pół) tekstowy przekaz. Warunkiem tu jest znajomość reguł technicznych oraz kognitywno-percepcyjnych podstaw w projektowaniu takiego komunikatu.

Podsumowanie i dalsze badania

Dzisiejsze inicjatywy ukierunkowane na rozwój społeczeństwa informacyjnego, takie jak upublicznianie danych i raportów publicznych, sieciowy dostęp do specjalistycznych baz danych, ruch na rzecz otwartych danych kształtują świadomość potrzeby ich logicznego i efektywnego prezentowania. Nadchodząca epoka *big data* wskazuje nowe sposoby zdobywania nowej wiedzy, oparte na analizie, przetwarzaniu i wizualizacji różnorodnych, nieustrukturalizowanych danych sieciowych. W świecie nauki w ostatniej dekadzie znajduje to odbicie w szybkim rozwoju interdyscyplinarnych badań nad wizualizacją informacji.

Potencjał tego nowego obszaru badań, a w szczególności opracowywane narzędzia i metody służące do wieloaspektowych analiz dostrzegają bibliolodzy i informatolodzy[26]. Po 2005 roku rozszerzyła się oferta globalnych naukowych baz danych.

[24] *Graphical abstracts*, Elsevier, [online] http://www.elsevier.com/authors/journal-authors/graphical-abstract [dostęp 30.06.2016]

[25] *Badania społeczeństwa informacyjnego*, Portal Pinterest, [online] https://pl.pinterest.com/veslavaosinska/badania-spo%C5%82ecze%C5%84stwa-informacyjnego [dostęp 30.06.2016].

[26] V. Osińska, *Mapowanie nauki i potencjał analityczny tego procesu*, „Zagadnienia Informacji Naukowej" 2010, 2(96), s. 41–51.

Nowe źródła i nowe algorytmy mapowania piśmiennictwa i innych form aktywności świata akademickiego to nowe możliwości i wyzwania dla informatologów. Kolekcje publikacji naukowych, indeksy cytowań, bazy danych bibliograficznych są przedmiotem badań bibliometrycznych, gdyż zawierają jednostki analityczne na różnym poziomie agregacji, takie jak: autorzy, dokumenty, czasopisma, instytuty i centra badawcze, środki finansowe na badania. Otwarty dostęp do pełnych tekstów, zapewniany przez coraz więcej wydawnictw, daje poszerzone możliwości badania dziedzinowych korpusów tekstowych, a z drugiej strony wymaga wydajnych metod automatycznej klasyfikacji i kategoryzacji dokumentów naukowych.

Wizualizacja wielkoskalowych danych pozwala na wychwycenie ich podobieństw, pogrupowanie tematyczne, dziedzinowe bądź instytucjonalne i tym samym określenie struktury obszaru badawczego, współczesnych kierunków i zakresów integracji multidyscyplinarnych. Na mapach nauki uczeni mogą zidentyfikować strukturę społeczności zajmującą się określonym obszarem badań, skupioną wokół konkretnego zagadnienia naukowego, a rozproszoną po całym globie. W ten sposób można wykryć wiodące ośrodki badawcze, kompetentnych naukowców, założycieli danej szkoły lub kierunku naukowego, najbardziej profesjonalne czasopisma. Wizualizacje mogą również pomóc w działaniach ewaluacyjnych w skali mikro np. w ocenie dorobku wybranych osób.

Autorzy przedstawili wyniki analiz metadanych polskich bibliotek cyfrowych w postaci graficznej, wybierając technikę wizualizacji stosownie do typu danych (liczba, tekst, daty) oraz poszukiwanych ukrytych podobieństw i korelacji. Obecnie udokumentowano ponad 200 mocno zróżnicowanych metod wizualizacji informacji[27]. Nie lada problem w ich wyborze i implementacji mają nie tylko początkujący użytkownicy danych, lecz także humaniści cyfrowi. Koncentrując się na tym wyzwaniu, autorzy zaprezentowali podstawowe kategorie metod *Infovis*. Uwzględniono tu cel końcowy i efekt: wykrycie zmian ilościowych na wykresie, pokazanie geograficznego rozrzutu danych, ujawnienie związków intelektualno-społecznych oraz efektywne komunikowanie idei eksperymentu. Jeśli odwołamy się do precyzji terminologicznej, ostatni przypadek nie jest wizualizacją informacji, lecz wizualizacją wiedzy, co często jest mylone ze względu na młodą dyscyplinę *Infovis* i brak jej bazy teoretycznej[28].

Jak zaznaczono wyżej metadane FBC wymagały wielu etapów czyszczenia. Dlatego to zadanie było najbardziej czasochłonne w realizacji całego eksperymentu dotyczącego analizy zmian strukturalnych w naukach humanistycznych. Jednak pozwoliło to autorom na wykrycie wielu nieprawidłowości w edycji danych i wytyczenie ścieżek naprawy złych praktyk w bibliotekach cyfrowych. W obecnej fazie prace są skoncentrowane na klasyfikacji dokumentów według dziedzin naukowych i odfiltrowanie dokumentów w zakresie nauk humanistyczno-społecznych. Przyszła baza danych zostanie poddana analizom pod względem współautorstwa, podejmowanej tematyki i zależności czasowych w trakcie powstawania publikacji. Reprezentacje graficzne zostaną porównane z ich odpowiednikami wygenerowanymi na podstawie danych z innych serwisów sieciowych.

[27] V. Osińska. *Wizualizacja informacji. Studium*…, s. 88–106.
[28] Ibidem, R. 2.

Podziękowania
Badania przeprowadzono w ramach grantu NCN 2013/11/B/HS2/03048. Autorzy wyrażają podziękowanie instytucjom, które udostępniły dane: Poznańskiemu Centrum Superkomputerowo-Sieciowemu, Ministerstwu Nauki i Szkolnictwa Wyższego oraz grupie CLARIN-PL za pomoc w tworzeniu klasyfikatora.

Bibliografia

Börner K., *The Atlas of Science,* USA 2010.
Chen Ch, *Information Visualization. Beyond the Horizon*, 2nd ed, London 2006.
Chen Ch., Kuljis J., *The rising landscape: a visual exploration of superstring revolutions in physics*, „Journal of the American Society for Information Science and Technology" 2003, 54(5), s. 435–446.
Garfield E., *Historiographic mapping of knowledge domains literature*, „Journal of Information Science" 2004, 30(2).
Garfield E., Pudovkin A. L., Istomin V. S., *Why do we need algorithmic historiography?*, „Journal of the American Society for Information Science and Technology" 2003, 54(5), s. 400–412.
Garfield E., *Scientography: Mapping the tracks of science*, „Current Contents: Social & Behavioural Sciences" 1994, 7(45).
Garfield E., Sher I. H., Torpie R. J., *The Use of Citation Data in Writing the History of Science*, Pensylvania, USA 1964.
Jóźwik K., Zwoliński Sz., *Myślenie wizualne w biznesie. Ty też potrafisz rysować*, Warszawa 2015
Lima M., *The Book of Trees. Visualizing Branches of Knowledge*, New York 2014.
Madsen R. E., Hansen L. K., Winther O., *Singular value decomposition and principal component analysis*, Raport techniczny 2004, [online] http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4000/pdf/imm4000.pdf [dostęp 30.06.2016].
Malak P., *Indeksowanie treści*, Warszawa 2012.
Malak P., Pawłowski A., *Ewaluacja skuteczności systemów wyszukiwania informacji. Od eksperymentu Cranfield do laboratoriów TREC i CLEF. Geneza, metody i wyniki*, „Toruńskie Studia Bibliologiczne" 2015, 1(14).
Miller G. A., *The magical numer seven, plus or minus two: some limits of our capacity for processing information*, „Psychological Review" 2001, t. 101, nr 2, s. 343–352.
Osińska V., Osiński G., Kwiatkowska A. B., *Visuallization in Learning: Perception, Aesthetics and Pragmatism*, [w:] *Maximiziną Coqnitive Learniną throuqh Knowledqe Visualization*, red. A. Ursyn, Hershey, PA 2015, r. 13.
Osińska V., *Mapowanie nauki i potencjał analityczny tego procesu*, „Zagadnienia Informacji Naukowej" 2010, 2(96), s. 41–51.
Osińska V., *Wizualizacja i wyszukiwanie dokumentów*, Warszawa 2010.
Osińska V., *Wizualizacja informacji. Studium informatologiczne*, Toruń 2016.
Osińska V., *Wizualizacja paradygmatów naukowych*, „Zagadnienia naukoznawstwa" 2012, nr 48, s. 205–220.
Stephen F., *Now you see it. Simple Visualization techniques and Quantitative Analysis*, CA, USA 2009.
Strang G., *Introduction to linear algebra*, 4th Edition, Cambridge, UK 2009, [online] http://math.mit.edu/~gs/linearalgebra [dostęp 30.06.2016].

Źródła elektroniczne

Badania społeczeństwa informacyjnego, Portal Pinterest, [online] https://pl.pinterest.com/veslavaosinska/badania-spo%C5%82ecze%C5%84stwa-informacyjnego/ [dostęp 30.06.2016].

Clarin PL – Polska część infrastruktury naukowej CLARIN ERIC, [online] http://ws.clarin-pl.eu/demo2/tager.shtml [dostęp 30.06.2016].

Federacja Bibliotek Cyfrowych, [online] http://fbc.pionier.net.pl/ [dostęp 30.06.2016].

GapMinder, [online] https://www.gapminder.org/ [dostęp 30.06.2016].

Garfield E., *Essays/Papers on Mapping the World of Science*, [online] http://garfield.library.upenn.edu/ mapping/mapping.html [dostęp 30.06.2016].

Graphical abstracts, Elsevier, [online] http://www.elsevier.com/authors/journal-authors/graphical-abstract [dostęp 30.06.2016].

Hyperbolic browser, WikiViz, [online] http://www.wikiviz.org/wiki/Hyperbolic_browser [dostęp 30.06.2016].

Kulczycki E., *Post na blogu*, [online] https://www.facebook.com/emanuelkulczycki?fref=ts [dostęp 30.06.2016].

Kulczycki E., *Warsztat badacza – Emanuel Kulczycki*, [online] http://ekulczycki.pl/ [dostęp 30.06.2016].

Leiden Manifest for Research Metrics, [online] http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351 [dostęp 30.06.2016].

Lengler R., Eppler M., *Visualisation methods Periodic Table*, 2007, [online] http://www.visual-literacy.org/periodic_table/ periodic_table.html [dostęp 30.07.2016].

Madsen R. E., Hansen L. K., Winther O., *Singular value decomposition and principal component analysis*, Raport techniczny 2004, [online] http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4000/pdf/imm4000.pdf [dostęp 30.06.2016].

Schabish J., *The Graphic Continuum*, [online] http://visual.ly/graphic-continuum [dostęp 30.06.2016].

Strang G., *Introduction to linear algebra*, 4th Edition 2009, [online] http://math.mit.edu/~gs/linearalgebra/ [dostęp 30.06.2016].

Walrus – Gallery: Visualization & Navigation, Center for Applied Internet Data Analysis, [online] https://www.caida.org/tools/visualization/walrus/gallery1 [dostęp 30.06.2016].

Weingart S., *Diagrams of knowledge*, Blog 2013, [online] http://www.scottbot.net/HIAL/?page_id=39166 [dostęp 30.06.2016].

Wikipedia. *Hasło Eugene Garfield*, [online] https://en.wikipedia.org/wiki/Eugene_Garfield [dostęp 30.06.2016].

Wizualizacja informacji, [online] http://www.wizualizacjainformacji.pl/onas.php [dostęp 30.07.2016].

The development of research on visualization of information

Abstract

The last decade has shown a rapid development of interdisciplinary research on visualization of information (Infovis). In bibliology, the first person to become interested in these methods was a pioneer of science-metrics, E. Garfield, who introduced the term "science-grams". In creation of "science-grams", or maps of science, are involved IT specialists and bibliologists, computer specialists and science specialists, as well as specialists on presentation of

information and graphic designers. This article aims at presenting the possibilities of using Infovis as a new field of research in information sciences through a series of examples of visualization for the collection of metadata, that come with bibliographic descriptions of documents available in Polish digital libraries. The results of the analyzes of metadata the authors presented in graphical form, choosing the technique of visualization according to the data type (number, text, date) and according to context analysis. The targets of data visualization: detection of quantitative changes on a graph, showing the geographical spread of the data, the disclosure of intellectual and social relationships, and effective communication of the idea of the experiment, are described in detail and illustrated.

Key words: data visualization, maps, information, visual science-metrics, Infovis

Veslava Osińska
University of Nicolaus Copernicus in Toruń
Institute of Science Information and Bibliology

Piotr Malak
Univeristy of Wrocław
Institute of Science Information and Library Science

Bożena Bednarek-Michalska
University Library of Nicolaus Copernicus University in Toruń
Certified curator