Badanie struktury i dynamiki zasobów cyfrowej wiedzy przy pomocy metod wizualizacji - projekt realizowany na UMK

Veslava Osińska[1] (wieo@umk.pl)**, Piotr Malak**[1](piomk@umk.pl)**, Bożena Bednarek-Michalska**[2](bozena@umk.pl)

[1]*Instytut Informacji Naukowej i Bibliologii, Uniwersytet Mikołaja Kopernika w Toruniu*
[2] *Biblioteka Uniwersytecka, Uniwersytet Mikołaja Kopernika w Toruniu*

Abstrakt. Zjawiskiem towarzyszącym cyfryzacji zasobów wiedzy naukowej jest marginalizowanie w mniejszym lub w większym stopniu meta-opisu danych strukturalnych. W wyniku tego mamy albo zaśmiecone metadane albo ich brak dla serii rekordów. Jest to główny problem, z którym zetknęli się autorzy przy realizacji projektu dotyczącego analizy dynamiki cyfrowej wiedzy w Polsce. Czyszczenie i grupowanie danych zostało wykonane na pierwszym etapie w środowiskach Python i R. Do wstępnych analiz i wypracowania dalszej strategii posługiwano się wizualizacją danych tekstowych za pomocą chmury słów. W charakterystykach dynamicznych wykorzystano konwersję danych tekstowych na format daty oraz metody statystyczne. Autorzy wprowadzają w problematykę przetwarzania metadanych pochodzących z bibliotek cyfrowych i nakreślają odpowiednie strategie postępowania.

Słowa kluczowe: biblioteki cyfrowe, wizualizacja wiedzy, mapy nauki, lingwistyka komputerowa.

Wprowadzenie

Biblioteki i repozytoria cyfrowe są obecnie solidnym źródłem aktualizowanej na bieżąco otwartej wiedzy dla naukowców. W zasobach tych można znaleźć zarówno treści specjalistyczne, jak i popularne, naświetlające problemy ale nie dotykające obszarów naukowych. Analizy, jak szybko i w jakim zakresie zachodzą zmiany tych treści, mogą dostarczyć ciekawych informacji o rozwoju współczesnej cyfrowej wiedzy, natomiast badania nad rozłożeniem tematyki pomogą opisać jej strukturę. Metodologia badawcza opiera się na naukometrii, zintegrowanej z zaawansowanymi algorytmami wizualizacji dużych zbiorów danych, gdzie przyjęło się wykorzystywać teorię grafów oraz/lub sztuczne sieci neuronowe. Ostatnie generują wynik w postaci mapy, odwzorowującej cechy danych wejściowych.

Wynikowe reprezentacje graficzne, zwane mapami wiedzy (albo równie często mapami nauki, mapami informacji w zależności od poruszanego tematu, ale również i od audytorium) umożliwiają eksplorację danych w nowy sposób, prowadzący do odkrywania nowych, istotnych zależności i zjawisk dotąd nieobserwowanych. Dlatego

wśród specjalistów w dziedzinie wizualizacji informacji często używa się metafory „odkrywanie nowej wiedzy" w procesie studiowania map wizualizacyjnych[1].

O projekcie

Problematykę wizualizacji w naukometrii podjęto w roku 2014 na Uniwersytecie Mikołaja Kopernika w ramach projektu finansowanego z funduszy NCN „Badanie struktury i dynamiki zasobów cyfrowej wiedzy". Badania mają na celu między innymi stworzenie serii funkcjonalnych map, obrazujących stan współczesnej wiedzy cyfrowej w Polskiej humanistyce. Powstałe wizualne struktury planuje się przeanalizować w zestawieniu z odpowiednikami odnoszącymi się do nauki światowej, a tworzonymi przez zagraniczne zespoły akademickie i publikowanymi na dedykowanych portalach, takim jak np. *Places & Spaces*[2]. W efekcie można będzie dostrzec zasadnicze różnice w tempie i kierunkach rozwoju polskiej i światowej humanistyki. Badania cyfrowych zasobów wiedzy naukowej dają podstawy do sformułowania następującego pytania badawczego i poszukiwania na nie odpowiedzi: w jakim stopniu polska humanistyka jest cyfrowa? Jednym z celów badania jest przedstawienie charakterystyki współpracy polskich uczonych w zakresie humanistyki, określającej strukturę społeczną danego obszaru badawczego.

W cyfrowych zasobach akademickich wykorzystywanymi jednostkami analizy stają się przede wszystkim metadane artykułów naukowych. Żeby wyłuskać z nich dane nadające się do wizualizacji, trzeba zastosować algorytmy przetwarzania i eksploracji tekstu oraz statystykę opisową. Jak pokazuje doświadczenie zagranicznych kolegów, najbardziej pracochłonnym etapem w pracach wizualizacyjnych jest skompletowanie i przetworzenie danych badawczych. Co więcej, czasem ich specyfika i uwarunkowania techniczne decydują o sukcesie eksperymentu, o tym, np., czy dane nadają się do wizualizacji, albo czy wyjściowe reprezentacje są materiałem merytorycznym i/lub rzetelnym. Niniejszy artykuł drąży problematykę doboru właściwych zbiorów danych do tego rodzaju prac i nakreśla pewną strategię postępowania.

---

[1] M. Lima, *The Book of Trees. Visualizing Branches of Knowledge*, Architectural Press, New York: Princeton 2014; K. Börner, *Everyone can map*, The MIT Press, USA: Cambridge 2014.

[2] *Places & Spaces: Mapping Science*. [data dostępu 28.01.2016]. Dostępny w: http://scimaps.org.

Grupa robocza składa się z trzech osób z przydziałem ściśle wyprofilowanych zadań. **Bożena Bednarek-Michalska** jest działaczką Ruchu Open Access i specjalizuje się w analizie i usprawnianiu funkcjonowania bibliotek cyfrowych oraz repozytoriów naukowych[3]. Dr **Piotr Malak** rozwija algorytmy eksploracji tekstu oraz lingwistyki komputerowej i jest odpowiedzialny za przetwarzanie danych. Kierownik projektu – dr **Veslava Osińska** wizualizuje dane, uprzednio je grupując według wspólnych cech ustalonych w wyniku analizy statystycznej danych. Opis prowadzonych prac badawczych oraz uzyskane wyniki są publikowane na dedykowanym portalu Wizualizacja Nauki: http:wizualizacjanauki.umk.pl.

Zasoby polskich bibliotek i wybranych repozytoriów cyfrowych są w Polsce udostępniane na ujednoliconej platformie sieciowej FBC (Federacja Bibliotek Cyfrowych)[4]. Obecnie oferuje ona otwarty dostęp do ponad 2,5 mln[5] jednostek zbiorów cyfrowych, która to kolekcja stale rośnie. Autorzy portalu usprawnili wyszukiwarkę publikacji poprzez włączenie operacji logicznych na polach opisu dokumentu. Lista wyników skojarzona jest z odnośnikami do strony internetowej biblioteki cyfrowej, która zawiera znaleziony obiekt cyfrowy. Kluczowe dla założeń projektu było pozyskanie oraz przeanalizowanie zawartości metadanych artykułów o profilu naukowym, bowiem w bibliotekach cyfrowych, składowane są również dokumenty nie mające nic wspólnego z kontekstem akademickim, np.: pisma urzędowe, ulotki i broszury reklamowe, spisy wystaw itp. FBC stosuje w opisach dokumentów standard Dublin Core, udostępniając 15 pól opisu w opcji wyszukiwania zaawansowanego: od tytułu, twórcy, tematu, opisu, wydawcy aż po źródło, język, zakres i prawa. Kolekcję metadanych do analiz projektowych udostępnili twórcy FBC z Poznańskiego Centrum Superkomputerowo-Sieciowego[6].

Kolejnym źródłem danych cyfrowych reprezentujących prace naukowe są repozytoria instytucji naukowych i badawczych. Część danych dostępnych w tym typie źródeł jest obecna na platformie FBC, jednakże pełniejszy zakres danych oferuje

[3] *Uwolnij naukę*, [data dostępu 28.01.2016]. Dostępny w: http://uwolnijnauke.pl/.
[4] *Federacja Bibliotek Cyfrowych*. [data dostępu 28.01.2016]. Dostępny w: http://fbc.pionier.net.pl/.
[5] Stan na styczeń 2016 r.
[6] *PCSS*, Poznań. [data dostępu 28.01.2016]. Dostępny w: http://www.man.poznan.pl/online/pl/.

agregator Centrum Otwartej Nauki (CEON)[7]. Z agregatora CEON pozyskaliśmy do tej pory 58 674 rekordów, natomiast są to wyłącznie dane dotyczące polskich publikacji naukowych. Analiza naukometryczna zasobów dostępnych w repozytoriach naukowych jest kolejnym etapem opisywanych badań.

Ponieważ jednym z elementów badań naukometrycznych podjętych w ramach omawianego projektu jest analiza współpracy między instytucjami oraz między pracownikami nauki polskiej, dane identyfikujące naukowców pracujących w polskich jednostkach badawczych i naukowych pozyskaliśmy z OPI[8].

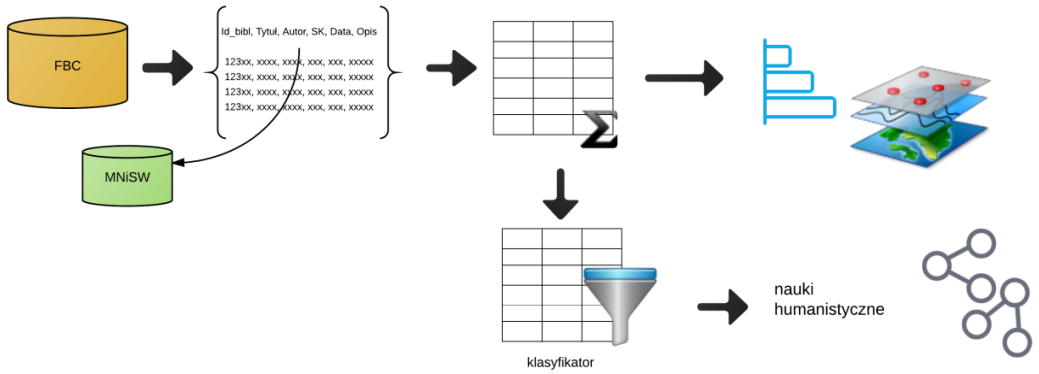
Zakres prac i analiz

Obecnie przeanalizowaliśmy najliczniejszy z wymienionych zasobów - zbiór FBC. Pierwszy etap prac dotyczył czyszczenia danych do postaci nadającej się do analiz statystycznych i kompatybilnej ze stosowanymi aplikacjami (rys. 1).

W trakcie początkowych analiz i opracowań odkryliśmy, że biblioteki cyfrowe nie stosują albo stosują w różnym stopniu dyscyplinę jednorodności przy wprowadzeniu danych w takich polach jak: opis, słowa kluczowe, data powstania czy typ obiektu cyfrowego, co bardzo utrudnia wnioskowanie na podstawie danych. Ze względu na wysoką wiarygodność i jakość danych o autorstwie postanowiono przeanalizować relacje zachodzące w ich obrębie. Dane jako informacyjnie jakościowe (*informative*), zostały wykorzystane do badań współautorstwa i liczebności grup badawczych[9] (rys. 2). Wyniki analizy wskazują, że reprezentacja prac naukowych współtworzonych przez kilku autorów jest bardzo duża w bibliotekach cyfrowych. Być może jest to tendencja w metodologii współczesnej pracy naukowej, która wynika z pracy zespołowej tak łatwej w środowisku elektronicznej komunikacji i szybkiej wymiany myśli? Z kolei analiza afiliacji autorów poszczególnych prac wykazała, że współpraca zespołowa odbywała się na poziomie różnych instytucji.

[7] *CEON. Agregator.* [data dostępu 28.01.2016]. Dostępny w: http://agregator.ceon.pl/
[8] *OPI – Ośrodek Przetwarzania Informacji*.[data dostępu 28.01.2016]. Dostępny w: http://www.opi.org.pl/
[9] V. Osińska, P. Malak, *Maps and Mapping in Scientometrics*, Proceedings of the Conference *Tools and Methods for Analysing the Scientific Literature and Readers*, Wrocław 2014 (in print).

Rysunek 1. Etapy przetwarzania i wizualizacji danych.



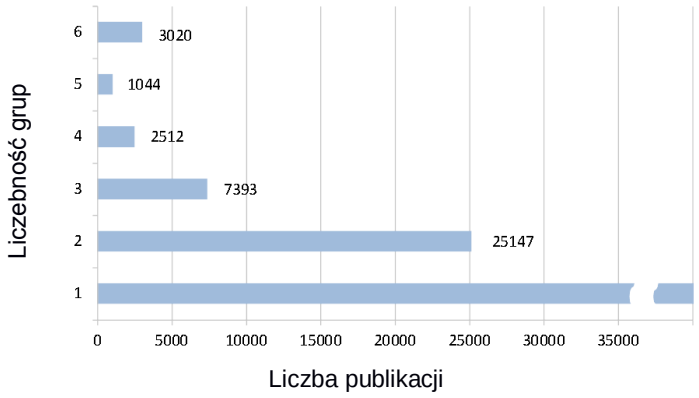
Źródło: opracowanie własne.

Korzystając z bazy danych pracowników naukowych w polskich instytucjach naukowo-badawczych, udostępnionych przez OPI, skorelowaliśmy dane dotyczące autorów z ich afiliacjami i ostatecznie stworzyliśmy mapę współpracy pomiędzy polskimi instytucjami naukowymi (rys. 3).

Ze względu na specyfikę systemu FBC, indeksującego metadane dokumentów z różnych dziedzin, w czym większość indeksowanych zasobów stanowią prace nienaukowe, trzeba było odfiltrować zbiory z zakresu humanistyki, co okazało się niełatwym zadaniem. Metadane nie dostarczają wskazówek wystarczających do wiarygodnego wyselekcjonowania dowolnej dziedziny naukowej. W pracach skoncentrowanych na dokładnym określeniu dziedziny analizowano takie pola jak: tytuł (*dc*:*Title*), przedmiot (*dc*:*Subject*) i opis (*dc*:*Description*). W przypadku dostępnych poprzez FBC prac naukowych tytuły najczęściej dotyczyły bardzo ściśle określonych tematów lub przedmiotów badań, na przykład materiałów i narzędzi w naukach inżynierskich, algorytmów w informatyce, epoki historycznej w naukach historycznych.

Sporo szumu informacyjnego wprowadzały czasopisma, których w polskich bibliotekach naukowych jest bardzo dużo. Spora część rekordów (60%-90% w zależności od biblioteki, 76% całego otrzymanego z FBC zbioru) zawierała opisy czasopism i gazet. W tym przypadku pole *dc:Title* było więc zupełnie nieprzydatne do analiz dziedzinowych, tym bardziej, że w niektórych przypadkach oprócz nazwy czasopisma zawierało także rok i numer wydania. Powtarzalność nazwy w wielu rekordach generuje redundancję danych, co ponownie dowodzi niereprezentatywności samego tytułu w zakresie wskazania dziedziny naukowej.

Rysunek 2. Liczba publikacji (poziomo) „wyprodukowana" w wieloosobowych zespołach (pionowo).



Źródło: opracowanie własne.

Pole temat (*dc:Subject*) najczęściej składało się z szerzej lub węziej określonych słów kluczowych albo tematu badań. Zdarzało się też odniesienie do autora dokumentu, lub kraju pochodzenia. Niespodziewanie spora liczba rekordów (ok. 20%) w zbiorze nie miała wypełnionego pola *dc:Subject*. Opis (*dc:Description*) również nie zaliczał się do konsekwentnie wypełnianych pól. Pole to było wypełnione tylko w około 20% z pozyskanych rekordów. Jedynie dla niewielkiej części zbioru (zaledwie około 5%) było to streszczenie artykułu naukowego w ujęciu klasycznym.

W końcowej ocenie autorów nasunęła się konkluzja o losowo dobieranych charakterystykach dokumentów, a w efekcie oceny całego uzyskanego zbioru metadanych o braku jakichkolwiek standardów i ustaleń, nawet w ramach jednej biblioteki. Problemy te wskazywały jasno, że automatyczna analiza danych pod kątem wyodrębnienia polskiej humanistyki naukowej powstałej po roku 1945 będzie bardzo trudna.

Problemy z jednolitym opisem metadanych

Ze wstępnej analizy wynikało zatem, iż z pól standardu Dublin Core najbardziej wartościowe dla opisywanych badań były pola: *dc:Subject, dc:Title, dc:Type, dc:Author* oraz *dc:Date*. Początkowe założenie, że zawartość pola *dc:type*, opisująca typ publikacji będzie przydatna przy filtrowaniu publikacji naukowych okazało się niestety trudne. Wynikało to z braku jednoznacznych standardów oraz niekontrolowanego wypełniania tego pola przez bibliotekarzy.

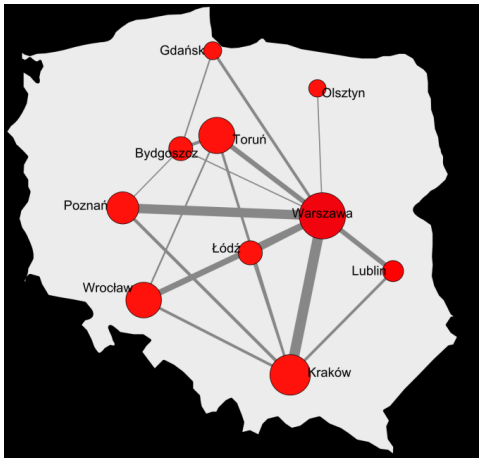
Bardzo często zamiast faktycznych typów dokumentów w polu tym pojawiały się inne treści. Były to między innymi:

- tytuł podany również jako typ,
- wymiary dokumentu fizycznego,
- autor dzieła,
- data publikacji dzieła.

Typy nie były robione wg żadnych norm polskich czy europejskich. W zdecydowanej większości polskich bibliotek cyfrowych, których zbiory dostępne są za pośrednictwem FBC można zauważyć problem braku standaryzacji opisów również na płaszczyźnie językowej, bardzo często bowiem pojawiał się dany typ publikacji ale pisany raz w liczbie pojedynczej, raz w liczbie mnogiej. Nie było to dużym problemem podczas automatycznej analizy danych, ale jednak nie świadczy zbyt dobrze o jakości procesu tworzenia opisów dokumentów cyfrowych. Kolejnym problemem napotkanym podczas analizy typów dokumentów był brak listy lub słownika limitujących liczbę typów. Powoduje to z jednej strony nadmierne rozdrobnienie typologii zbiorów, z drugiej zbytnią generalizację typów. Ta wielorakość i bałagan spowodował, że badacze pochylili się nad ujednoliceniem typologii.

Podobne problemy z niejednorodnością opisu występowały w polu *dc:date*, oznaczającym datę publikacji dzieła. Można było w nich wykryć różnorodne standardy oraz nieścisłości.

Rysunek 3. Współpraca polskich ośrodków naukowo-badawczych w zakresie publikowania na podstawie danych bibliotek cyfrowych. Szerokość pasma pokazuje intensywność współpracy (tj. liczbę wspólnych publikacji), natomiast rozmiar kółka – ilość ośrodków.



Źródło: opracowanie własne.

Oprócz różnej kolejności części opisujących rok, miesiąc i dzień, pojawiają się różne separatory w formacie daty, ale również inne wyrażenia, jak: „po YYYY", „przed", „post ante", „ok.", „ca ...". To z kolei powodowało problemy z eliminacją utworów opublikowanych przed, przyjętym za graniczny, rokiem 1945.

Wypracowane strategie badawcze

Dla skutecznej analizy oraz wiarygodności wyciąganych na jej podstawie wniosków w automatycznym badaniu tekstów, a tym są metadane, konieczne jest ujednolicenie posiadanych danych. W obliczu wielkiej swobody bibliotekarzy we wprowadzaniu metadanych do systemu oraz braku szczegółowych ustaleń normalizacyjnych dotyczących zawartości poszczególnych pól standardu Dublin Core konieczne było przygotowanie zestawu algorytmów, które ujednoliciłyby zawartość interesujących nas pól metadanych. W przypadku dat publikacji (*dc:Date*) wykorzystano wyrażenia regularne (*regular expressions*) do konwersji różnych form zapisu daty na jeden wspólny format. W przypadku zaś pól opisujących typ dokumentu (*dc:Type*) podjęto prace nad przygotowaniem roboczego słownika hierarchicznego terminów i na jego podstawie automatycznie ujednolici się opisy. Ma to doprowadzić do uporządkowania typów dokumentów charakterystycznych dla obszaru nauki i wyeliminowania błędnych określeń.

Do wstępnych analiz i zaplanowania kolejnych kroków w pracach posłużono się wizualizacją słów kluczowych za pomocą chmury słów. Jako źródła danych wykorzystane zostały kolekcje słów pól *dc:subject* i *dc:description* (rys. 4). Wykonano kilka wersji wizualnych reprezentacji i przedstawiono do oceny bibliotekarzowi, mającemu wieloletnią praktykę w rzeczowym opracowaniu dokumentów. Ekspert ostatecznie wybrał wizualizację pokazana na rysunku 4, wymieniając kilka dominujących obszarów tematycznych, min.: historia Polski XX stulecia, czasopisma z zakresu bibliologii, język informacyjno-wyszukiwawczy KABA, problematyki bibliotek cyfrowych i in.

Pomimo, iż taka interpretacja powstała na skutek działań *ad hoc*, można było wywnioskować o znaczącej przewadze dokumentów z zakresu nauk humanistycznych. Nie rozwiązało to jednak problemu odfiltrowania tych dokumentów.

Rysunek 4. Chmura korpusu słów dla pól Subjects i Description

Źródło: opracowanie własne.

Do selekcji artykułów z zakresu nauk humanistyczno-społecznych wykorzystany został automatyczny klasyfikator zbudowany na bazie narzędzi CLARIN-PL[10]. Należy tu nadmienić, iż pozostała część zbioru zawierała dokumenty odnoszące się do nauk przyrodniczych, technicznych, rolniczych, medycznych i ekonomicznych. Wartości pól *dc:Title, dc*:*Subject, dc:Description* były szczegółowo analizowane w kontekście dyscyplinarności. Tworzyły one zbiór, bardzo zróżnicowany pod względem składni, struktury i semantyki. Dlatego autorzy postanowili wspomóc się tezaurusem (wielo)dziedzinowym, który powiązałby reprezentatywne terminy wymienionych pól zgodnie z semantyką słownika języka polskiego.

Tym samym pełniłyby one rolę haseł przedmiotowych, a powstałe relacje generowałyby taksonomię danej dyscypliny. Przejrzano wiele tezaurusów dostępnych zarówno w Europie jak i w Polsce, ale nie odpowiadały one na zapotrzebowanie projektu. Zdecydowano się zatem na prostsze rozwiązanie, wyznaczenie docelowej struktury dziedzinowej, adekwatnej do roboczej kolekcji FBC, z która przyszło się badaczom zmierzyć. Autorzy zadecydowali, iż najprostszym rozwiązaniem będzie wykorzystanie klasyfikacji paneli dziedzinowych, wypracowaną przez Narodowe Centrum Nauki. Panel humanistyczno-społeczny (HS) składa się z trzech głównych kategorii (HS1-HS3) oraz drugiego poziomu rozszczepionego na dyscypliny naukowe opisane za pomocą wyrażeń (rys. 5).

[10] CLARIN-PL. Europejska sieć współpracy naukowej. Grupa badaczy pracująca pod kierunkiem dr inż. Macieja Piaseckiego podjęła się zadania stworzenia i implementacji klasyfikatora przy użyciu słownika semantycznego Słowosieć. http://clarin-pl.eu/pl/strona-glowna/

Rysunek 5. Zastosowany schemat kategorii dziedzinowych na podstawie paneli NCN.

HS1 : filozofia i etyka
HS1 : teologia i religioznawstwo
HS2 : literaturoznawstwo
HS2 : językoznawstwo
HS2 : kulturoznawstwo
HS2 : bibliologia i informatologia
HS2 : sztuka i architektura i historia sztuki
HS3 : historia
HS3 : archeologia
HS3 : etnologia, etnografia i antropologia kulturowa
HS3 : archiwistyka i dokumentologia
HS6 : psychologia
HS6 : pedagogika
HS6 : socjologia

Źródło: opracowanie własne.

Podsumowanie i wnioski

Realizatorzy projektu zamierzają zbadać, jaka jest cyfrowa reprezentacja wiedzy naukowej i edukacyjnej w polskich repozytoriach i bibliotekach cyfrowych, dlatego prowadzą analizę *text-* i *datamining* w odniesieniu do nauk humanistycznych i społecznych. Platforma polskich bibliotek cyfrowych FBC stosuje standardy opisu bibliograficznego *Dublin Core*. Dlatego na obecnym etapie baza danych FBC została wybrana jako zbiór modelowy w procesach przetwarzania różnych metadanych.

Podczas realizacji wyznaczonych celów badawczych, autorzy spotkali się z licznymi problemami, bezpośrednio dotyczących uzyskanych danych. Do nich można zaliczyć brak dyscypliny przy wprowadzaniu metadanych (na przykład w polach data, typ), typowe błędy ludzkie (literówki) lub po prostu brak opisów. Dodatkowo udostępnieniu metadanych z repozytoriów cyfrowych towarzyszyły wątpliwości natury prawnej. Na przykład: czy zebrane dane publiczne, odpowiednio przetworzone, można później publikować, jako wyniki swoich badań? Czy można je udostępniać innym podmiotom?

W obecnej fazie prace są skoncentrowane na stworzenie klasyfikatora, pozwalającego na odfiltrowanie utworów naukowych w zakresie nauk humanistyczno-społecznych. Takim sposobem przygotowana baza danych zostanie następnie poddana analizom pod względem autorów i wzajemnej współpracy, tematyki, zależności

czasowych powstawania publikacji. Reprezentacje graficzne zostaną porównane z ich odpowiednikami wygenerowanymi w oparciu o dane z innych serwisów sieciowych.

Artykuł przybliża wypracowane przez zespół techniczne i wizualne strategie, służące do zniwelowania problemów z zapisem danych w bibliotekach cyfrowych. Pokonanie tych trudności nie oznacza jednak, iż nie ma potrzeby ujednoliconych i ustandaryzowania danych, uskuteczniających operacje z zakresu Humanistyki Cyfrowej.

Podziękowania

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

Bibliografia

Bednarek-Michalska B., Repozytoria surowych danych - dlaczego biblioteki powinny je znać?, ”*Biuletyn EBIB 2012”*, nr 8 (135), s. 1-8.

Bednarek-Michalska B., Kujawsko Pomorska Biblioteka Cyfrowa a standardy, „*Biuletyn EBIB” 2006*, nr 4/(74): http://www.ebib.info/2006/74/michalska.php [dostęp online: 28.01.2016].

Börner K., *Everyone can map*, The MIT Press, USA:Cambridge 2014.

Chen Ch. *Information Visualization. Beyond the Horizon*, Springer, London 2006.

Lima M., *The Book of Trees. Visualizing Branches of Knowledge*, Architectural Press, New York: Princeton 2014

Malak P., *Indeksowanie treści*, SBP, 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, nr 1 (14).

Nahotko M., Stare i nowe standardy opisu dokumentów elektronicznych.*Biuletyn EBIB*, 2002 nr 4, Standardy i organizacja. [data dostępu 28.01.2016]. Dostępny w: http://www.ebib.pl/2002/33/nahotko.php.

Potęga J., *Metadane w polskich bibliotekach cyfrowych*. Warszawa BN 2008. [data dostępu 28.01.2016]. Dostępny w: www.bn.org.pl/download/document/1260454699.

Osińska V., Wizualizacja paradygmatów w nauce, „*Zagadnienia naukoznawstwa*” 2012, 48 (193).

Osińska V., Malak P., *Maps and Mapping in Scientometrics*, Proceedings of the Conference *Tools and Methods for Analysing the Scientific Literature and Readers*, Wrocław 2014 (in print).

Osińska V., Rozwój metod mapowania domen naukowych i potencjał analityczny w nim zawarty, „*Zagadnienia Informacji Naukowej*” 2010, 2(96).

Osinska V., Visual Analysis of Classification Scheme, “*Knowledge Organisation*” 2010, 37(4).

Werla M. Metadane dokumentów w bibliotekach cyfrowych. PCSS Poznań 2009. [data dostępu 28.01.2016]. Dostępny w: *http://lib.psnc.pl/Content/284/CPI-Werla.pdf.*

The project at the NCU: Information Visualization methods in digital knowledge structure and dynamics study

Abstract. Parallel to the growth of digital knowledge we can observe low importance of data structure. As a result, the metadata are cluttered or have empty values in a pieces of records. This is the main problem experienced by the authors in the project focused on analysing the dynamics of digital knowledge in Poland. Cleaning and grouping of data by use Python and R has been done in the first stage. The authors use cloud tag technique for preliminary analysis and to develop further strategies. Text data conversion into date format as well as the statistical methods were used to describe dynamic characteristics of published documents. The authors introduce the issues of metadata processing in digital libraries and outline an appropriate strategies to do it.

Keywords: digital libraries, knowledge visualization, maps of science, computational linguistics, text mining.