Jakość wybranych elementów metadanych stosowanych w polskich bibliotekach i repozytoriach cyfrowych[1]

Piotr Malak[2]
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytet Wrocławski

Veslava Osińska
Instytut Informacji Naukowej i Bibliologii
Uniwersytet Mikołaja Kopernika w Toruniu

Bożena Bednarek-Michalska
Biblioteka Uniwersytecka
Uniwersytet Mikołaja Kopernika w Toruniu

Abstrakt

Cel/Teza: Niniejszy artykuł prezentuje wyniki badań nad oceną jakości metadanych w polskich bibliotekach cyfrowych oraz możliwości wykorzystania tych danych do masowego przetwarzania zawartości cyfrowych repozytoriów oraz wyodrębnienia współczesnych prac naukowych.
Koncepcja/Metody badań: Badania zostały przeprowadzone hybrydowo – poprzez automatyczną analizę metodami inżynierii lingwistycznej oraz analizę ekspercką.
Wyniki i wnioski: W wyniku przeprowadzonych badań zidentyfikowano szereg nieścisłości i nieprawidłowości występujących w metadanych dla dokumentów w polskich bibliotekach cyfrowych. Wskazano również propozycje poprawy tego stanu.
Zastosowanie praktyczne: Wnioski płynące z badań mogą przyczynić się do znaczącej poprawy jakości metadanych i umożliwienia wykorzystania metadanych jako wartościowego materiału badawczego.
Oryginalność/Wartość poznawcza: Według najlepszej wiedzy autorów badania tego typu, zarówno w zakresie tematyki jaki i skali, nie były dotychczas prowadzone w Polsce.

Słowa kluczowe:

Analiza metadanych. Biblioteki cyfrowe. Jakość metadanych. Dublin Core.

Otrzymany: 9 czerwca 2016. Zrecenzowany: 22 lipca 2016 oraz 5 września 2016.
Poprawiony: 17 sierpnia 2016 oraz 19 stycznia 2017. Zaakceptowany: 10 marca 2017.

1. Wprowadzenie

W ramach projektu „Badanie struktury i dynamiki cyfrowych zasobów wiedzy za pomocą metod wizualizacji" (ang. *Information Visualization methods in digital knowledge structure*

[1] Badania przeprowadzono w ramach grantu NCN 2013/11/B/HS2/03048. Badanie przeprowadzono przy współpracy z realizatorami projektu *Polska część infrastruktury naukowej CLARIN ERIC* CLARIN PL.
[2] Kontakt z autorami artykułu za pośrednictwem dr. Piotra Malaka: piotr.malak@uwr.edu.pl

and dynamics study), finansowanego w latach 2014–2017 przez Narodowe Centrum Nauki, a realizowanego na Uniwersytecie Mikołaja Kopernika (UMK), prowadzony jest szereg badań nad dokumentami cyfrowymi i ich opisami bibliograficznymi jako zasobami wiedzy[3]. Jednym z ważnych tematów badawczych podjętych w ramach tego grantu jest **analiza i ocena jakości metadanych** w polskich bibliotekach cyfrowych. Badanie to prowadzono w celu sprawdzenia możliwości wykorzystania metadanych do masowego przetwarzania zawartości cyfrowych repozytoriów i wyodrębnienia na ich podstawie współczesnych prac naukowych. W wyniku przeprowadzonych badań i analiz dokonano oceny jakości metadanych w polskich bibliotekach cyfrowych, wskazano grupy błędów występujących w opisach oraz zgłoszono postulaty działań zmierzających do poprawy jakości metadanych. Celem niniejszego artykułu jest zaprezentowanie wyników badań nad jakością metadanych w polskich bibliotekach cyfrowych, wniosków płynących z analizy oraz propozycji poprawy bieżącego stanu.

Badaniami opisanymi w niniejszym artykule objęto wszystkie biblioteki cyfrowe w Polsce, w których do opisu dokumentów stosuje się schemat metadanych Dublin Core[4]. Na potrzeby omawianych badań przyjęto następujące założenia:

(1) analizie poddane są metadane używane w opisach publikacji dostępnych w polskich bibliotekach cyfrowych;
(2) prace opublikowane przed rokiem 1945 nie spełniają warunku współczesności;
(3) obiekty oznaczone jako należące do domeny publicznej nie spełniają wymogu współczesności. Założenie to wynika z faktu, że do domeny publicznej trafiają w większości prace starsze, niepodlegające już ochronie prawnej, 70 lat od śmierci autora lub dacie publikacji (w przypadku czasopism).

Spełnienie powyższych założeń związane jest z celami grantu badawczego, jest również wystarczające dla wstępnej selekcji i kwalifikacji obiektów do analizy jakości metadanych na potrzeby niniejszego artykułu.

W omówionych w artykule badaniach jakości metadanych przyjęto następujące hipotezy:

(1) Duża część metadanych stosowanych w polskich bibliotekach cyfrowych nie spełnia kryteriów jakości określonych na potrzeby omawianych badań.
(2) Analiza wartości elementów opisu Dublin Core: dc:rights, dc:type, dc:date umożliwia wyselekcjonowanie współczesnych zasobów naukowych z całości zasobów polskich bibliotek cyfrowych, w których stosowany jest ten standard.
(3) Opisy wydawnictw ciągłych, oznaczonych jako czasopisma, gazety, itp. nie zawierają informacji wystarczających do oceny czy wydawnictwo to ma charakter naukowy, ani tym bardziej do określenia tematyki badawczej poruszanej w publikowanych artykułach. W metadanych opisujących czasopisma i gazety brak jest odniesienia do ich zawartości (zob. Tab. 1). Są to zazwyczaj opisy wydawnictwa zbiorowego, czasem uzupełnione o spisy treści występujące jako samodzielne dokumenty.

Zgodnie z hipotezą 2. do analiz wybrano zawartość następujących elementów metadanych standardu *Dublin Core*:

[3] O badaniach związanych z grantem zob. m.in. V. Osińska, P. Malak (2016a; 2016b); V. Osińska, P. Malak, B. Bednarek-Michalska (2016).

[4] Opisany m.in. w PN-ISO 15836:2006 Informacja i dokumentacja – – Zestaw elementów metadanych Dublin Core.

– *dc:date* – pole zawierające najczęściej datę, na dowolnym poziomie szczegółowości, powstania lub udostępnienia obiektu, norma *The Dublin Core Metadata Element Set* (DCMES) zaleca stosowanie zapisu dat w postaci zgodnej z normą ISO 8601[5], czyli w postaci: RRRR-MM-DD. Jak podaje M. Nahotko (2000), możliwe jest użycie innej formy zapisu dat, ale powinna ona zostać jednoznacznie zidentyfikowana. Dobrym przykładem takiego podejścia do notowania dat jest opracowanie przygotowane przez Bibliotekę Uniwersytecką we Wrocławiu (BUWr) (*Interpretacja schematu Dublin Core*, 2006).
– *dc:type* – pole opisujące rodzaj dokumentu pod względem kategorii lub gatunku. Standard *DCMI Metadata Terms*[6] zaleca korzystanie ze słownictwa kontrolowanego, zalecanego w dokumencie *DCMI Type Vocalbulary*. Są to następujące terminy: Collection, Dataset, Event, Image, Interactive Resource, Service, Software, Sound, Text[7]. Również opracowanie BUWr, w oparciu o Polską normę PN-92 N-01227, podaje słownik kontrolowany typów zasobów. Na dziewięć wyróżnionych w nim kategorii typów formy dokumentu podane są aż trzy synonimy.
– *dc:rights* – informacja o prawach własności intelektualnej, prawach autorskich lub prawach własności.

Tab. 1. Przykładowe metadane czasopisma dostępnego w bibliotece cyfrowej [data dostępu 20 kwietnia 2016]

Przegląd Biblioteczny 2000, z. 4.
Dublin Core wer.1.1 :
Tytuł : Przegląd Biblioteczny 2000, z. 4.
Wydawca : H. Dobrzycki ; Wydawnictwo SBP ; Zakład Narodowy im. Ossolińskich
Miejsce wydania : Warszawa
Współtwórca : Sordylowa, Barbara. Red.
Instytucja sprawcza : Polska Akademia Nauk. Biblioteka w Warszawie ; Stowarzyszenie Bibliotekarzy Polskich
Data wydania : 2000
Typ zasobu : czasopismo
Format : image/x.djvu
Identyfikator zasobu : oai:bbc.uw.edu.pl:112
Język : pol

Źródło: http://bbc.uw.edu.pl/publication/148

1.1. Przyjęte kryteria jakości metadanych

Metadane w systemach informacyjno-wyszukiwawczych powinny z założenia spełniać kryteria jakości, czy to wynikające z zastosowanego schematu, czy też z wewnętrznych regulacji instytucji prowadzącej dany system. Do podstawowych funkcji metadanych opisowych należy dostarczenie informacji jednoznacznie identyfikujących konkretny obiekt w zbiorze. Aby

[5] Date and Time Formats – W3CDT, https://www.w3.org/TR/NOTE-datetime
[6] *DCMI Metadata Terms*, http://dublincore.org/documents/dcmi-terms/
[7] Za *DCMI Type Vocabulary*, http://dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/

ta funkcja była realizowana właściwie, w przypadku dokumentów tekstowych w bibliotekach cyfrowych zazwyczaj wystarczy prawidłowo podać dane identyfikacyjne dokumentu, jak *tytuł*, *autor*, *identyfikator w systemie*. Metadane schematu Dublin Core dostarczają również informacji przybliżających treść opisywanego dokumentu. W tym przypadku standardy jakości wyznaczone są przez przyjęty system opisu. Na potrzeby omawianych badań przyjęto kryteria oceny jakości metadanych wynikające z założeń badawczych całego projektu grantowego. Podstawowym kryterium było więc dostarczenie informacji, które pozwolą efektywnie wyszukać w dostępnym zbiorze publikacji dokumenty o zadanych parametrach. Parametrami tymi były: współczesność dokumentu (*dc:date*), określona na mniej niż 70 lat od daty powstania utworu (*dc:rights*), naukowy charakter treści dokumentu (*dc:type*) oraz tematyka związana z zakresem nauk humanistycznych. Pole *dc:rights* pozwoliło wyeliminować wszystkie obiekty z zapisem *domena publiczna* jako stare, niewspółczesne – zawartość tego pola może być traktowana komplementarnie do zawartości pola dc:date, szczególnie w sytuacjach, gdy data publikacji nie została podana. Ze względu na przyjęte założenia, istotnymi elementami metadanych są zapisy w polach *dc:type*, *dc:rights*, *dc:date*. Proces automatycznego wyszukiwania i klasyfikowania informacji na podstawie zawartości tych pól wymaga, aby dane w nich zawarte pozwalały efektywnie wyodrębnić dokumenty z poszukiwanej kategorii. Aby spełnić te wymagania opisy powinny spełniać następujące kryteria:

– spełniać zalecenia norm;
– jednoznacznie opisywać dokument;
– zachować konsekwencję w stosowanych wyrażeniach.

Zachowanie wymienionych kryteriów, w odniesieniu do każdego pola opisu w metadanych, niezależnie od przyjętego systemu, leży w interesie każdej biblioteki cyfrowej, ponieważ m.in. ułatwia wymianę danych między systemami, a także efektywne wyszukiwanie informacji o dokumentach we własnym systemie. Należy tutaj podkreślić, że we współczesnych bibliotekach cyfrowych zazwyczaj obiektami przeszukiwanymi nie są pełne teksty dokumentów, lecz właśnie metadane opisujące te dokumenty (Baca, 2016). Metadane są również uniwersalnym nośnikiem informacji w procesach wymiany danych pomiędzy różnymi bibliotekami, np. Federacja Bibliotek Cyfrowych (FBC)[8] prezentuje wyłącznie metadane pozyskiwane z polskich bibliotek cyfrowych. Stąd dbałość o jakość i poprawność danych wprowadzanych do poszczególnych pól opisu powinna być istotnym elementem polityki udostępniania zasobów.

1.2. Przegląd piśmiennictwa

Problematyka jakości metadanych wpisuje się w ogólny nurt rozważań nad jakością bibliotek cyfrowych, nabierając szczególnego znaczenia w kontekście szerokiego udostępniania i promowania danych z zakresu dziedzictwa kulturowego, w tym m. in. poprzez propagację w Europeanie. W literaturze światowej i polskiej problem metadanych oraz ich jakości pojawia się regularnie. Tym bardziej dziwi fakt, że wiele z podnoszonych przy tej okazji postulatów albo nie budzi zainteresowania twórców bibliotek cyfrowych, albo z innych powodów nie jest uwzględniane w ich doskonaleniu. Poniżej na podstawie wybranych prac krótko scharakteryzowane zostały przykłady problemów podejmowanych w polskich publikacjach

[8] http://fbc.pionier.net.pl/

poświęconych temu tematowi. Ponieważ problem dotyczy polskich bibliotek cyfrowych, przegląd piśmiennictwa został ograniczony do prac polskich badaczy i praktyków.

A. Domagalska (2006) przeanalizowała dostępne w latach 2002–2006 trzy rozwiązania standaryzacyjne z zakresu tworzenia bibliotek cyfrowych (NISO 2004, Strategia Europejska I2010, Strategia Zespołu ds. Standardów dla Bibliotek Naukowych) konstatując, że w żadnym z nich nie poświęcono wcale uwagi kwestiom oceny jakości bibliotek cyfrowych, a tym samym pominięto również kwestie jakości metadanych. W pracy podniesiony został problem wagi metadanych w procesie wyszukiwania treści cyfrowych (w tym zdigitalizowanych) oraz wpływu jakości tych metadanych na efektywność wyszukiwawczą danego systemu, co z kolei przekłada się na poziom satysfakcji użytkowników. Podano tam również trzy typy metadanych: opisowe, administracyjne oraz strukturalne. Autorka pracy podniosła również kwestię konieczności stworzenia, przynajmniej na poziomie instytucji, systemu autorytatywnej kontroli metadanych oraz zwróciła uwagę na fakt, że za błędne metadane odpowiada człowiek – podczas ich tworzenia, lub maszyna – w przypadku automatycznego generowania na podstawie dostarczonych danych.

W pracy zbiorowej pod redakcją G. Płoszajskiego (2008) również rozważano temat kategorii metadanych. Wskazano podstawy prawne regulujące ich definicje oraz różnice pomiędzy metadanymi opisującymi cyfrowe kopie obiektów fizycznych a tymi, które opisują dokumenty utworzone oryginalnie jako cyfrowe. W pracy tej podniesiono również problem metadanych dla archiwów, bibliotek i muzeów oraz rozróżniono „obiekt oryginalny i odpowiadający mu obiekt cyfrowy", a także wskazano na możliwe różnice w zawartości metadanych opisujących oba typy obiektów. W publikacji tej Europeana jest wyróżniona ze względu na stosowanie metadanych opisowych jako metadanych wyszukiwawczych. Stawia się tam również postulat przyjęcia konkretnej listy typów dokumentów. Największy jednak nacisk położony został na omówienie metadanych technicznych i administracyjnych. Publikacja stanowi zbiór opisów dobrych praktyk związanych z tworzeniem i udostępnianiem kolekcji cyfrowych, popartych przykładami istniejących wzorcowych standardów oraz wdrożeń. Nie dokonuje jednakże oceny bieżącego stanu metadanych w polskich bibliotekach cyfrowych.

Praktycznym problemom wykorzystania metadanych z polskich bibliotek cyfrowych poświęcona została praca M. Werli (2010). Jej autor zaprezentował doświadczenia Poznańskiego Centrum Superkomputerowo-Sieciowego (PCSS) – twórcy FBC, w wykorzystaniu metadanych do tworzenia graficznych interfejsów eksploracji zasobów bibliotek cyfrowych. Autor opisał proces analizy metadanych pod kątem określenia daty oraz miejsca wydania publikacji, czyli w zakresie częściowo pokrywającym się z tematyką badań opisywanych w niniejszym artykule. M. Werla opowiada się za postulatem stosowania słowników lub kartotek haseł wzorcowych przy tworzeniu metadanych w celu osiągnięcia odpowiedniego poziomu ich interoperacyjności. Twierdził, że w przypadku danych typu data, gdzie trudno wykorzystać słowniki zamknięte, należy wdrożyć normalizację ich zapisu, jako przykład podając notację RRRR-MM-DD. Autor postulował również takie zaprojektowanie schematu metadanych, aby możliwe było automatyczne wyodrębnienie określeń przestrzennych czy czasowych. Na podstawie analiz danych przeprowadzonych na potrzeby niniejszego artykułu, można niestety stwierdzić, że stan tego typu metadanych w ciągu sześciu lat od ukazania się pracy Werli nie uległ zmianie. FBC wprowadziła system automatycznego poprawiania i uzupełniania metadanych przesyłanych przez biblioteki cyfrowe, m.in. w zakresie ujednolicania zapisów dat, ale system taki nie zastąpi rzetelnego wprowadzania danych przez biblioteki.

W odczuciu autorów niniejszego artykułu smutną wymowę tego braku poprawy wzmacnia fakt, że polskie biblioteki cyfrowe udostępniają swoje dane za pośrednictwem FBC, gdzie M. Werla pełni funkcję Kierownika Działu Bibliotek Cyfrowych i Platform Wiedzy, a mimo to nie wdrażają wysuwanych przez niego propozycji poprawy jakości metadanych.

Z kolei artykuł M. Nahotko (2010) ma charakter przeglądowy, opisuje badania krajowe oraz międzynarodowe w zakresie automatycznego generowania metadanych oraz przybliża aplikacje stosowane do pracy z metadanymi do 2010 r. Autor duży nacisk położył na wskazanie możliwości wykorzystania nowoczesnych metod indeksowania treści do generowania metadanych opisujących dokumenty.

Warto też wspomnieć o serwisie *Digitalizacja.pl*, poświęconym problemom digitalizacji zbiorów bibliotecznych, muzealnych i archiwalnych. Znajdują się tam wartościowe poradniki oraz dyskusje, w tym dotyczące metadanych i ich jakości. Należy również odwołać się do powstałego w 2012 r. raportu pt. *Metadane, zagadnienia słowników kontrolowanych* (2012), który porusza problematykę rozwiązań dotyczących digitalizacji obiektów muzealnych i ich udostępniania, omawiając również kwestie metadanych oraz dostarczając propozycje rozwiązań w tym zakresie.

Temat metadanych dyskutowany jest również w opracowaniach poświęconych generalnej ocenie jakości bibliotek cyfrowych. Analizy te jednak dotyczą zazwyczaj bibliotek cyfrowych jako serwisów internetowych, ich budowy, dostępności oraz funkcjonowania (Kazan, Skubała, 2008; Potęga, 2009; Głowacka, 2011; Derfert-Wolf, 2011; Żernicka, 2014). Metadane są tam wymieniane jako jeden ze składników oceny jakości biblioteki cyfrowej jako całości, bez poświęcania uwagi jakości samych metadanych. J. Potęga (2009) podaje dodatkowo wyniki analiz ilościowojakościowych wykorzystania w opisach bibliograficznych pól schematu Dublin Core oraz zawartości tych pól. Z wyjątkiem pracy K. Żernickiej (2014), pozostałe tu wymienione opisują stan do 2011 r.

2. Metodyka badań

Przedmiotem badań były metadane obiektów udostępnianych w bibliotekach cyfrowych bez względu na oryginalną formę wydawniczą. Uwzględniono dokumenty upowszechniane od 2007 r., od kiedy funkcjonuje Federacja Bibliotek Cyfrowych, która pełni funkcje agregatora metadanych z polskich bibliotek cyfrowych i repozytoriów naukowych. Warto tutaj nadmienić, że w ramach projektu „Badanie struktury i dynamiki cyfrowych zasobów wiedzy za pomocą metod wizualizacji" badania prowadzone są nie tylko na danych udostępnianych za pośrednictwem FBC, ale także na zasobach udostępnianych przez inne instytucje – Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego (ICM UW), które prowadzi agregator polskich repozytoriów naukowych i Repozytorium Centrum Otwartej Nauki (CEON)[9] oraz przez Ośrodek Przetwarzania Informacji (OPI)[10], zbierający dane osobowe o naukowcach polskich[11].

[9] http://agregator.ceon.pl

[10] http://www.opi.org.pl

[11] Ze względu na odmienny charakter danych z OPI, nie przeprowadzono ich analizy jakościowej. Dane te nie zostaną również omówione w niniejszym artykule, ponieważ nie mieszczą się w jego zakresie tematycznym.

Biblioteki i repozytoria cyfrowe gromadzą różne zasoby, zarówno ze względu na tematykę, jak i formę publikacji, w tym także prace naukowe, które są głównym przedmiotem badań wspomnianego projektu badawczego. Różnorodność zgromadzonych w nich materiałów wymagała więc znalezienia sposobu wyodrębniania spośród nich współczesnych publikacji naukowych, do czego planowano wykorzystać metadane trzech wyróżnionych typów. Metadane zarówno w repozytoriach, jak i w bibliotekach cyfrowych prezentowane są w tym samym schemacie, tj. Dublin Core, co pozwala na zastosowanie do ich oceny jednakowych metod badawczych.

Badanie metadanych można przeprowadzić na wiele sposobów, korzystając z metod automatycznych albo za pomocą analiz jakościowych wykonywanych przez eksperta. Oba podejścia mają swoje zalety oraz ograniczenia. Wykorzystanie technik automatycznych jest mniej pracochłonne w dłuższej skali czasowej, jednakże problemem może być niejednolity sposób zapisu metadanych, co utrudnia ich automatyczne przetwarzanie do celów badawczych. Spełnienie warunku jednolitości zapisu, jako silnie zależnego od człowieka, niestety nie zawsze jest możliwe. Różna pragmatyka indeksowania i staranność wprowadzania metadanych przez różne instytucje i indeksatorów skutkuje znacznym zróżnicowaniem metadanych opisujących te same własności dokumentów. Z kolei analizy eksperckie są zdecydowanie bardziej czaso- i pracochłonne, i nie gwarantują sukcesu ze względu na rozmiary zbiorów danych do opracowania. Dlatego na potrzeby opisywanych badań zastosowano podejście zintegrowane. Dane badawcze były pobierane oraz przetwarzane automatycznie, zaś analizy wykonywano hybrydowo, z wykorzystaniem automatycznych metod NLP (ang. *Natural Language Processing*), wizualizacji informacji oraz analiz eksperckich.

W procesie badawczym zaplanowano następujące kroki:

(1) pobieranie wszystkich metadanych z wymienionych wcześniej platform agregujących zasoby cyfrowe;
(2) przetwarzanie metadanych i ocena ich jakości.

3. Pobieranie metadanych

Do realizacji etapu pierwszego zdecydowano się wykorzystać istniejące rozwiązania informatyczne takie jak eksport z bazy danych oraz automatyczne pobieranie danych. Metadane z FBC zostały udostępnione dzięki uprzejmości pracowników PCSS, zaś dane z Agregatora CEON pobrano za pomocą protokołu OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), co pozwala na automatyczne uaktualnianie tych danych w przyszłości. Ogółem, na początku projektu pobrano 1 714 571 rekordów ze 101 bibliotek i repozytoriów cyfrowych dostępnych za pośrednictwem FBC (w tym 87 bibliotek cyfrowych) oraz 53 899 rekordów dostępnych w Agregatorze CEON[12].

Dane zostały poddane operacjom usunięcia duplikatów, konwersji zapisu kodowego na UTF-8 oraz wyodrębnienia informacji z pól wskazanych we wcześniejszej części niniejszego artykułu.

[12] W chwili oddania artykułu do druku Agregator CEON rejestrował 71 606 dokumentów.

4. Analiza metadanych

W niniejszym podrozdziale zostanie zaprezentowany opis stanu metadanych w polach odpowiadających trzem wybranym elementom standardu Dublin Core, tj. *dc:type*, *dc:rights* oraz *dc:date*.

4.1. Element TYP

Ze względu na fakt, że do badań wykorzystano jedynie metadane dokumentów cyfrowych, bez ich pełnej treści, wstępną klasyfikację zasobów na naukowe oraz nienaukowe przeprowadzono na podstawie zawartości pola opisu typ dokumentu (*dc:type*). W tym polu pojawiają się określenia wskazujące czy dany dokument jest np. dysertacją naukową, artykułem, pracą dyplomową czy monografią.

W metadanych polskich bibliotek cyfrowych można zaobserwować brak jakiejkolwiek normalizacji, co powoduje, że w poszczególnych polach opisów pojawiają się hasła niejednolite w treści (wykorzystanie różnych synonimów tego samego hasła podstawowego) i formie gramatycznej (niekonsekwentne stosowanie w opisach liczby pojedynczej i mnogiej). Dotyczy to również pola *dc:type*, także w przypadku określeń wskazujących na potencjalną pracę naukową. W polu tym pojawiają się na przykład następujące wartości:

- *e-rozprawa habilitacyjna*,
- *e-rozprawa habilitacyjna PŁ*,
- *habilitacja*,
- *habilitacje*,
- *rozprawa habilitacyjna*,
- *dysertacja*,
- *dysertacje*.

Sama niekonsekwencja w stosowaniu liczby pojedynczej i mnogiej nie stanowi dużego problemu podczas automatycznej analizy danych, ponieważ stosunkowo łatwo można było takie opisy ujednolicić, np. za pomocą lematyzacji[13]. Wymaga to jednak dodatkowego nakładu pracy analitycznej oraz odpowiedniego zaprojektowania systemu przetwarzania danych. Należy natomiast podkreślić, że występowanie tego zjawiska w metadanych z tej samej biblioteki cyfrowej świadczy o nieprzestrzeganiu lub braku standardów jakości przy wprowadzaniu opisów zasobów.

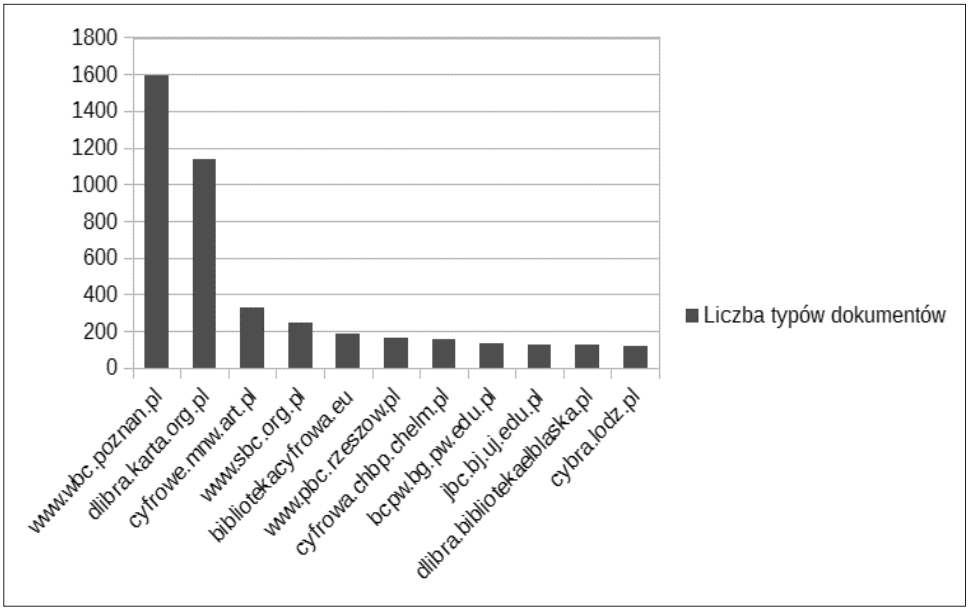
Ze względu na wymienione niekonsekwencje w opisach typów dokumentów stosowanych przez różne biblioteki cyfrowe problematyczne okazało się sformułowanie zapytania wyszukiwawczego, które uwzględniłoby wszystkie potencjalne typy prac naukowych stosowane przez różne biblioteki cyfrowe. Zastany brak jednolitości i konsekwencji w stosowaniu terminów w polach opisu schematu Dublin Core wymusza zastosowanie kilku zapytań wyszukiwawczych dla tego samego typu dokumentu (dotyczy również np. takich samych praw dostępu, itp.).

Dodatkowym utrudnieniem w klasyfikowaniu dokumentów na podstawie metadanych były umieszczane w polu *dc:type* wartości całkowicie nieprawidłowe, jak na przykład: tytuł, wymiary fizyczne oryginału, data publikacji, autor.

[13] Lematyzacja – proces wyznaczania podstawowej formy gramatycznej, na przykład mianownika liczby pojedynczej, dla wyrazów w tekście.

W wielu przypadkach zaobserwowano nadreprezentację typów dokumentów. Sporo bibliotek, niebędących bibliotekami specjalistycznymi, zbytnio rozdrobniło dane, co także utrudnia klasyfikację dokumentów wyłącznie na podstawie ich typów. Na 87 analizowanych zbiorów danych z bibliotek cyfrowych, które były dostępne na platformie FBC na początku badań, aż 11 rejestrowało powyżej 100 typów dokumentów włączając w tę liczbę typy opisane zarówno w liczbie pojedynczej, jak i w mnogiej oraz określenia podane w językach obcych. Te ostatnie traktowane były jako wyrażenia wielowyrazowe, zawierające termin w języku polskim i obcym, przy czym często w ramach opisów jednej biblioteki występowało kilka form zapisu. Zestawienie tych bibliotek oraz liczby rejestrowanych w nich typów dokumentów prezentuje rysunek 1.

Rys 1. Biblioteki rejestrujące powyżej stu typów dokumentów



Źródło: opracowanie własne.

W Wielkopolskiej Bibliotece Cyfrowej[14] odnotowano rekordową liczbę 1594 stosowanych określeń dla typów dokumentów (wliczając w to tłumaczenia nazw typów na język angielski oraz niemiecki), w tym m.in. różne nazwy o tym samym znaczeniu. Różnorodność określeń typów nie dotyczy zatem tylko odrębnych bibliotek, ale występuje także w ramach zasobu pochodzącego z jednej biblioteki cyfrowej, co bardzo utrudnia badanie zasobów, ich grupowanie, czy ocenę jakości metadanych. W tabeli 2 zaprezentowano zestawienie wartości pola *dc:type* dla typu dokumentu: *artykuł* stosowanego w różnych bibliotekach cyfrowych (po przecinku podane są zawartości powielonego pola *dc:type*: tłumaczenia lub synonimy). Należy przy okazji zauważyć, że samo określenie *artykuł* jest niejednoznaczne i nie pozwala określić czy opisuje pracę naukową czy publicystyczną.

[14] http://www.wbc.poznan.pl

Tab. 2. Zestawienie wartości pola dc:type dla typu dokumentu *artykuł* w różnych bibliotekach cyfrowych

article
article, artykuł, press article
artikel, artykuł, article
artikel, artykuł, press article
artykuł konferencyjny
artykuł naukowy
artykuł z czasopism ogólnopolskich
artykuł z czasopisma
artykuł, article, art
art z czasopisma
artykuł, articles
artykuł, postprint
artykuły
artykuły historyczno-prawne, zeszyt naukowy
artykuły z czasopism ogólnopolskich
artykuły, czasopisma
artykuły, maszynopisy
artykuły, opracowania historyczne
artykuły, polemiki
artykuły, recenzje
e-artykuł
fragment artykułu

Dużą różnorodność form typu publikacji może zobrazować przykład Bałtyckiej Biblioteki Cyfrowej[15]. W metadanych z tej biblioteki odnotowano 185 form zapisu typów dokumentów, wliczając w to połączone w listę zapisy z pól *dc:type* dla języków polskiego, niemieckiego oraz angielskiego :

- *magazin, czasopismo, journal*;
- *zeitschrift, czasopismo, journal*;
- *zeitschriften, czasopismo, journal*;
- *zeitschrift, gazeta, journal*;
- *czasopismo, journal*;
- *zeitschrift, czasopismo, periodical*.

O ile dla języka angielskiego można stwierdzić pewną jednolitość formy, to dla języków niemieckiego i polskiego widać dużą swobodę w doborze terminów opisujących typ dokumentu. W zapisie oryginalnym metadanych pobranych z FBC przykładowy rekord z BBC ma następujący zapis:

<dc:type xml:lang="de"><![CDATA[magazin]></dc:type>
<dc:type xml:lang="pl"><![CDATA[czasopismo]></dc:type>
<dc:type xml:lang="en"><![CDATA[journal]></dc:type>

Bezpośrednio na stronie Bałtyckiej Biblioteki Cyfrowej, w indeksie *Typ zasobu* można znaleźć m. in. hasła w liczbie pojedynczej i mnogiej, tzw. literówki w hasłach, itp. Zostały one zaprezentowane na rysunku 2.

[15] http://bibliotekacyfrowa.eu

Rys. 2. Indeks Typ zasobu, Bałtycka Biblioteka Cyfrowa, http://bibliotekacyfrowa.eu/dlibra/keywordindex?dirids=1&attId=9

Należy jednak podkreślić, że podobna sytuacja braku kontroli i konsekwencji w stosowaniu nazw typów dokumentów panuje w zdecydowanej większości bibliotek cyfrowych dostępnych poprzez platformę FBC. Wyniki wstępnego etapu badań, polegającego na analizach danych zawartych w opisach cyfrowych dokumentów, pozwalają wysnuć pesymistyczny wniosek, że typy dokumentów w polskich bibliotekach cyfrowych opisywane są w wielu przypadkach z pominięciem jakichkolwiek norm, czy to polskich (np. PN-92-N-01227) czy międzynarodowych (np. ANSI/NISO Z39.85–2012 The Dublin Core Metadata Element Set).

Nawet w sytuacji, gdy opisy typów dokumentów są ujednolicone i stosowane konsekwentnie, nie ułatwia to wyodrębnienia prac naukowych. W zestawieniu prezentującym szczegółową listę typów dokumentów stosowanych w metadanych z Kujawsko-Pomorskiej Biblioteki Cyfrowej[16] (KPBC), niewiele typów można zaliczyć do opisujących publikacje naukowe (wyróżnione w zestawieniu). Poważny problem interpretacyjny powstaje przy typie *książka* oraz *artykuł*, ponieważ pierwsze określenie może oznaczać zarówno monografię naukową, jak i beletrystykę, zaś drugie artykuł naukowy lub publicystyczny. Wartości

[16] http://kpbc.umk.pl. W KPBC wdrożono restrykcyjne zasady stosowania nazw typów, ograniczając tym samym ich redundancję, por. (Derfert-Wolf, 2016).

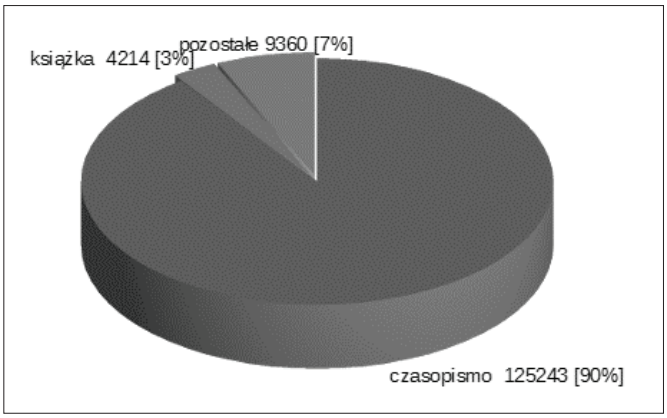
liczbowe umieszczone przy nazwach typów to łączne liczby udostępnionych publikacji zawierających w opisie podaną wartość atrybutu *dc:type*.

(1) czasopismo 125170
(2) książka 4214
(3) grafika 1420
(4) ekslibris 1035
(5) mapa 749
(6) druk muzyczny 593
(7) materiał aktowy 590
(8) pocztówka 548
(9) artykuł 543
(10) odbitka 499
(11) afisz 409
(12) malarstwo 404
(13) broszura 400
(14) fotografia 304
(15) rysunek 270
(16) nadbitka 188
(17) starodruk 180
(18) zaproszenie 147
(19) rękopis muzyczny 112
(20) inkunabuł 111
(21) rękopis 94
(22) preprint 94
(23) druk 87
(24) nagranie dźwiękowe 67
(25) przewodnik 36
(26) prezentacja 36
(27) jednodniówka 32
(28) katalog wystawy 28
(29) ulotka 26
(30) odezwa 25
(31) informator 22
(32) pamiętnik 21
(33) dyplom 21
(34) kopia rękopiśmienna druku muzycznego 20
(35) program 16
(36) list 16
(37) statut 14
(38) druk reklamowy 14
(39) program teatralny 13
(40) postprint 13
(41) plakat 12
(42) obwieszczenie 11
(43) śpiewnik 10
(44) maszynopis powielony 9
(45) atlas 9
(46) album 9
(47) wystawa 8
(48) raport 8
(49) film dokumentalny 8
(50) maszynopis 7
(51) kalendarz 7
(52) teka 6
(53) skrypt 5
(54) reprint 5
(55) wydawnictwo okolicznościowe 4
(56) partytura 4
(57) atlas starodruczny 4
(58) ulotka wyborcza 3
(59) postinkunabuł 3
(60) plan miasta 3
(61) ebook 3
(62) cennik 3
(63) legitymacja 2
(64) księga pamiątkowa 2
(65) znaczek okolicznościowy 1
(66) widok 1
(67) telegram 1
(68) tekst do mapy 1
(69) skład osobowy 1
(70) praca magisterska 1
(71) plan twierdzy 1
(72) pieśń 1
(73) partytura chóralna 1
(74) nekrolog 1
(75) mikrofilm 1
(76) komiks 1
(77) hasło słownikowe 1
(78) fotokopia 1
(79) film 1

Zaprezentowana różnorodność typów oraz problemy z jednoznacznym wskazaniem prac naukowych tylko na podstawie typu publikacji podanego w opisach dokumentów są wspólne dla wszystkich polskich bibliotek cyfrowych.

Na rysunku 3 przedstawiono udział dokumentów określanych jako książka, czasopismo lub za pomocą nazw innych typów w zasobach KPBC. Podobny rozkład typów dokumentów stwierdzono dla większości polskich bibliotek cyfrowych.

Rys. 3. Ilościowy rozkład typów dokumentów w zasobach KPBC, dane z 2 lutego 2016, źródło: http://kpbc.umk.pl/dlibra/pubstats



Przeważający typ zasobów w polskich bibliotekach cyfrowych stanowią wydawnictwa ciągłe, opisane zazwyczaj jako *czasopisma*. Na podstawie praktyki oraz analizy dat powstania zasobów można stwierdzić, że są to przeważnie dokumenty z domeny publicznej, czyli, zgodnie z założeniem opisywanych badań, niespełniające wymogu współczesności. Ponadto wiele bibliotek – w tym także KPBC – udostępnia czasopisma naukowe zdigitalizowane w całości, w zeszytach, bez rozbicia na poszczególne artykuły, co nie ułatwia analiz i selekcji danych. Ze względu na brak opisów indywidualnych artykułów, prace zawarte w tak zdigitalizowanych periodykach nie są dostępne do automatycznego przetwarzania tylko na podstawie metadanych. W celu ich analizy konieczne jest wyodrębnianie poszczególnych jednostek ze zdigitalizowanego zeszytu, rocznika, itp. Wymagałoby to użycia narzędzi rozpoznawania granic artykułów oraz wskazania ich części nagłówkowej (tytuł, autorzy, słowa kluczowe, itd.). W ramach omawianych badań tego rodzaju działań nie przeprowadzano.

Odmienną zawartość pod względem typów oraz wieku dokumentów reprezentują cyfrowe repozytoria naukowe. Zawierają one tylko współczesne prace naukowe. Z powodu małej liczby tego typu kolekcji cyfrowych w Polsce, w porównaniu do bibliotek cyfrowych, nie można było badań zawęzić jedynie do zasobów udostępnianych w repozytoriach. Ponadto analiza metadanych z bibliotek cyfrowych pozwala na wypracowanie bardziej uniwersalnych metod selekcji i analizy danych.

Ostatecznie, w wyniku analiz eksperckich, przeprowadzonych na materiale przygotowanym w procesie automatycznego przetwarzania i grupowania typów dokumentów, występujących w metadanych polskich bibliotek cyfrowych, wyodrębniono zbiór około 400 typów, które z dużym prawdopodobieństwem wskazują na pracę naukową. Poniżej zaprezentowany został fragment zestawienia (opracowanie własne na podstawie analizy i grupowania metadanych z polskich bibliotek cyfrowych):

– *art z czasopisma*
– *artykuł konferencyjny*

- *artykuły historyczno-prawne, zeszyt naukowy*
- *book, peerreviewed*
- *czasopismo naukowe polskie*
- *czasopismo, artykuł*
- *diplomarbeit, praca dyplomowa, diploma paper*
- *doktorat*
- *proceedings, e-lecture, audiovisual document*
- *dokument elektroniczny, dokument naukowo-dydaktyczny*
- *dokument elektroniczny, rozprawa habilitacyjna*
- *dokument naukowo-dydaktyczny*
- *dokument piśmienniczy, zeszyty naukowe politechnika łódzka, zeszyty naukowe pł*
- *e-książka, rozprawa habilitacyjna*
- *e-zeszyty naukowe pł, dokument elektroniczny, electronic resource, e-tul scientific bulletins*
- *fragment artykułu*
- *habilitacja*
- *hasło słownikowe*
- *komentarz do artykułu*
- *książka, rozprawa habilitacyjna*
- *książka, rozprawa habilitacyjna, book, dissertation.*

4.2. Element – PRAWA

Poprawne i prawidłowe rozpoznanie oraz oznaczenie praw do udostępnianego cyfrowo utworu jest niezwykle istotne. Pozwala użytkownikom rozpoznać zakres w jakim mogą z danego utworu korzystać (Bednarek-Michalska, 2014), a w przypadku badań przeprowadzanych na dużych zbiorach danych, za pomocą technik automatycznych, pozwala dokonać selekcję zasobów według kryteriów dostępności, ale także np. wieku publikacji. W trakcie opisywanych badań poddano zatem analizie również zawartość pola *dc:rights*. Jak wspomniano wcześniej, dla potrzeb projektu badawczego, którego celem było zbadanie struktury i dynamiki polskich naukowych zasobów cyfrowych przyjęto założenie, że obiekty oznaczone jako należące do domeny publicznej nie spełniają wymogu współczesności. Ocena współczesności dokumentu na podstawie zawartości pola *dc:rights* jest komplementarna w stosunku do oceny na podstawie zawartości pola *dc:date* – pozwala z jednej strony zweryfikować poprawność decyzji podjętej na podstawie podanej daty publikacji, a z drugiej strony pozwala sklasyfikować dokument, dla którego nie podano daty publikacji. Dzięki takiemu założeniu możliwe byłoby ignorowanie podczas badań wszystkich utworów trafiających do domeny publicznej (DP) z racji ich wieku.

W wyniku analizy zawartości pól *dc:rights* w danych z polskich bibliotek cyfrowych okazało się, że hipotezy tej nie można jednoznacznie uznać za potwierdzoną, ponieważ w większości bibliotek cyfrowych nieprawidłowo oraz niekonsekwentnie oznaczany jest status prawny utworu. Bardzo często w polu *dc:rights* spotykane są nieprawidłowości zarówno w treści, jak i formie wpisu. Analiza zawartości pól *dc:rights* wykazuje, że w większości zasobów polskich bibliotek cyfrowych, podobnie jak w przypadku pól *dc:type* oraz *dc:date*, panuje wielka dowolność oraz brak jakiejkolwiek kontroli treści i formy wpisów.

W wyszukiwarce FBC na zapytanie o *domenę publiczną* w polu PRAWA uzyskano w marcu 2015 r. liczbę 114 985 dokumentów. Jednakże na podstawie analizy zawartości zbiorów większości polskich bibliotek cyfrowych można stwierdzić, że było to zdecydowanie za mało (Derfert-Wolf, 2011). Do domeny publicznej powinno być przypisanych, zgodnie z ich faktycznym stanem prawnym, znacznie więcej zasobów z 2.5 mln, jakie były w badanym okresie udostępniane przez FBC.

W odpowiedzi na uwagę, że wyszukiwarka FBC może działać niepoprawnie, PCSS przekazało informację, że narzędzie to jest ciągle rozwijane i jeszcze nie obsługuje wielu typów zapytań. Twórcy FBC obiecali, że wprowadzą poprawki w systemie wyszukiwania. W styczniu 2016 r., po wdrożeniu przez PCSS poprawek w funkcjonowaniu wyszukiwarki FBC (m.in. wynikających ze zgłoszonych przez autorów niniejszego artykułu sugestii), ponownie zadano zapytanie: *dc_rights:(domena publiczna)* i uzyskano w odpowiedzi 1 206 874 dokumenty. Wartość ta wydaje się być zbliżona do faktycznej liczby tego typu dokumentów, jednakże nadal niedoszacowana, prawdopodobnie z powodu omówionych nieprawidłowości występujących w metadanych przekazywanych przez poszczególne polskie biblioteki cyfrowe.

Ewaluacje metadanych, przeprowadzone na potrzeby omawianych badań, dla zasobów bibliotek cyfrowych (z wyłączeniem repozytoriów) dostępnych w agregatorze FBC na początku 2015 r. wykazały, że na 1 650 075 rekordów w 1 429 267 (86% zasobów) z nich jest wypełnione pole *dc:rights*. W przypadku 762 588 rekordów (46% zbioru) pole to ma wartość *domena publiczna* w różnej, ale wciąż rozpoznawalnej jednoznacznie formie. Aż 13 spośród 87 analizowanych bibliotek cyfrowych nie wypełniało pola *dc:rights* w swoich metadanych. Natomiast łącznie w metadanych 35 bibliotek nie pojawiło się określenie „domena" w polu *dc:rights*.

Analiza zawartości pola wskazującego na prawa własności w bibliotekach cyfrowych dostępnych przez FBC wykazała aż 224 wzorce stosowane dla oznaczenia dzieł z domeny publicznej. Oprócz zalecanego określenia *domena publiczna*, można spotkać również m.in. wpisy:

– *domena publiczna (public domain),*
– *domena publiczna (public domain) / dla wszystkich bez ograniczeń,*
– *dostępne publicznie bez ograniczeń – domena publiczna.*

Niestety, większość wpisów obarczona była błędami. Należały do nich wskazywanie okresu powstania publikacji (np. XIII w., XVII–XVIII w.) lub daty śmierci autora czy też tłumacza, z zachowaniem różnych zasad stosowania znaków interpunkcyjnych, np.:

– *Domena Publiczna – Adam Mickiewicz zm. 1855*
– *Domena Publiczna – Adam Mickiewicz zm. 1855.*
– *Domena Publiczna – Adam Mickiewicz, zm. 1855*

Kolejnymi błędami występującymi we wpisach w polu *dc:rights* były tzw. literówki, np. Ulrich i Urlich (Leon Ulrich), czy wręcz błędy w dacie śmierci:

Domena Publiczna – Bolesław Leśmian zm. 1937
Domena Publiczna – Bolesław Leśmian zm. 1938

Na podstawie analizy wpisów w polu *dc:rights* w metadanych zbiorów cyfrowych można stwierdzić, że niektóre instytucje udostępniające zasoby cyfrowe rzetelnie wypełniają zawartość tego pola. W zakresie poprawności treści w opisach wyróżniają się np. zbiory Polskiej Akademii Nauk (PAN), dostępne na platformie Repozytorium Cyfrowego Instytutów

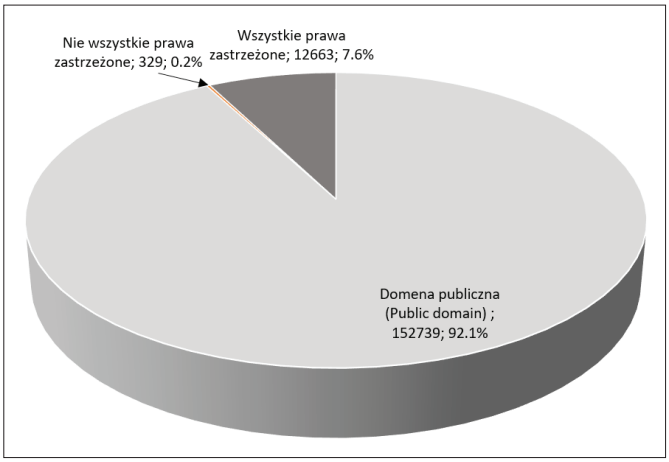
Naukowych (RCIN)[17]. Większość obiektów udostępnianych przez PAN poprzez platformę RCIN stanowią współczesne prace naukowe, zazwyczaj oznaczone jako chronione prawem autorskim: „copyright".

Kolejny przykład instytucji konsekwentnie i poprawnie oznaczającej prawa do obiektów stanowi Kujawsko-Pomorska Biblioteka Cyfrowa (KPBC). Zgodnie z wewnętrznymi wytycznymi KPBC każdy obiekt cyfrowy musi mieć pole „prawa" wypełnione wg przyjętego jednego wzorca, zaś bibliotekarze zostali w tym zakresie przeszkoleni[18]. Niżej przedstawiono przykład opisu dla czasopisma naukowego, które jest własnością UMK i wszelkie prawa zostały przez poszczególnych autorów przeniesione na uczelnię:

Tytuł: Acta Universitatis Nicolai Copernici. Nauki Matematyczno-Przyrodnicze. Geografia
Prawa: Wszystkie prawa zastrzeżone
Właściciel praw: Uniwersytet Mikołaja Kopernika w Toruniu
Licencja: Licencja UMK
Prawa dostępu: Dla wszystkich w zakresie dozwolonego użytku

Rysunek 4 prezentuje udział dokumentów o różnych prawach dostępu w zbiorach KPBC. Zaprezentowane proporcje są typowe dla polskich bibliotek cyfrowych (z wyłączeniem repozytoriów).

Rys. 4. Rozkład wartości pola PRAWA w KPBC.
Dane z 2 lutego 2016, źródło: http://kpbc.umk.pl/dlibra/pubstats



Analiza pola PRAWA pokazuje, że nie można bezkrytycznie wykorzystywać tych danych do selekcji dokumentów współczesnych. Przyczyną tego stanu jest niewypełnianie pola *dc:rights* przez wiele instytucji, lub też brak kontroli słownictwa używanego w opisach w przypadku większości bibliotek, które podawały status prawny publikacji. Opisy praw do obiektu mogą zostać wykorzystane w automatycznym przetwarzaniu metadanych pod warunkiem odpowiedniego ich wstępnego przetworzenia, np. poprzez uwzględnienie różnych potencjalnych forma zapisu oznaczającego domenę publiczną.

[17] http://rcin.org.pl
[18] Źródło: informacja od dyrekcji KPBC.

4.3. Element – DATA

Podobnie jak w przypadku pól *dc:type* i *dc:rights* problemy z niejednorodnością opisu występowały również w zasobach tej samej biblioteki cyfrowej w polu *dc:date* oznaczającym datę lub okres czasu związany ze zdarzeniem w cyklu istnienia zasobu. Można wskazać zastosowane różnorodne standardy zapisu daty, nie zawsze zgodne z zaleceniami DCMES, oraz nieścisłości w oznaczaniu dat publikacji. Sam zapis daty także przybierał rozmaite formy:

- RRRR-MM-DD
- DD-MM-RRRR
- RRRR.MM.DD
- DD.MM.RRRR

gdzie RRRRR oznacza rok w postaci pełnej (czterocyfrowej), MM – miesiąc, zaś DD – dzień.

Oprócz różnej kolejności elementów opisujących części daty, pojawiają się różne znaki oddzielające te części (np.: ‚.', ‚-'). W polach *dc:date* występują też inne sformułowania, nie zawsze konsekwentnie używane w obrębie metadanych z jednej biblioteki, jak:

- „po RRRR"
- „przed"
- „post ante"
- „ok."
- pierwsza połowa ... wieku
- „ca..."

Ponieważ jednak sam rok jako ciąg czterech następujących po sobie cyfr łatwo poddaje się automatycznemu rozpoznawaniu, stosunkowo prosto udało się z pól *dc:date* pozyskać roczne daty publikacji dokumentów. Zawartość tego właśnie pola stała się głównym wyznacznikiem współczesności publikacji w dalszych etapach badań.

5. Analiza ekspercka danych

W celu weryfikacji skuteczności zastosowanych w projekcie metod automatycznego przetwarzania i analizy statystycznej metadanych zdecydowano się przeprowadzić analizę ekspercką zasobów wybranych bibliotek. Decyzja ta wynikała również z faktu stosowania różnych zasad zapisu metadanych w poszczególnych bibliotekach cyfrowych, z czym wiązała się konieczność zastosowania bardzo ogólnych zasad wyboru i klasyfikacji dokumentów podczas etapu analizy automatycznej. Podczas analizy należało uwzględnić sytuację wydzielania, przez niektóre biblioteki cyfrowe, prac naukowych jako odrębnej kolekcji. W takich przypadkach podana również była liczba tych prac. Upowszechnienie tej praktyki podniosłoby wartość metadanych jako surowych danych badawczych oraz ułatwiłoby dalsze automatyczne analizy, m. in. naukometryczne.

Jako przykład wydzielonej kolekcji o charakterze naukowym może posłużyć zasób KPBC zatytułowany „Materiały dydaktyczne". Zgodnie z opisem: *Kolekcja zawiera artykuły, monografie i podręczniki akademickie wydane lokalnie po 1945 r. oraz opracowania historyczne niezbędne dla nauki i dydaktyki. Opiera się o prace pracowników nauki i wydawców naszego regionu, którzy zdecydowali się powierzyć nam prawa do ich udostępnienia. Liczba publikacji w kolekcji: 10946. Liczba publikacji w kolekcji i we wszystkich podkolekcjach: 11038* (KPBC, 2016).

W przypadku braku wydzielonych kolekcji o charakterze naukowym, konieczne było przeprowadzenie bardziej szczegółowych i wielopoziomowych analiz. Zastosowano metodę analizy jakościowo-ilościowej i podjęto następujące kroki, analogiczne do etapów przetwarzania i analizy automatycznej:

(1) przeglądanie danych statystycznych udostępnionych przez biblioteki cyfrowe w celu ustalenia:
   a. typów zasobów (analiza pola opisu TYP z metadanych);
   b. praw do zasobów (analiza pola PRAWA przeprowadzona w celu ustalenia liczebności zasobów w domenie publicznej);
(2) jeśli statystyki dostępne na stronach poszczególnych bibliotek nie podawały tych informacji korzystano z zaawansowanego wyszukiwania i przeszukiwano wskazane wyżej pola opisu;
(3) jeśli wyszukanie zaawansowane nie przynosiło oczekiwanych rezultatów, przeglądano dostępne kolekcje w celu określenia liczebności publikacji współczesnych (zasoby dydaktyczne, artykuły, doktoraty oraz starych (dziedzictwo kulturowe). Niekiedy pomocny był własny podział kolekcji wdrożony przez poszczególne biblioteki cyfrowe, nazwy kolekcji często sugerują, która część zawiera jakie publikacje. Wybór kolekcji powoduje wyświetlenie informacji o jej liczebności.

W tabeli 3 przedstawiono wyniki wybranych biblioteki poddanych badaniom i problemy, jakie rodzi podobna analiza. Do próby badawczej wybrano tylko te biblioteki, które mają większą liczbę zasobów. Wykluczono biblioteki, w których działanie wyszukiwarki wskazywało na ewidentne błędy techniczne.

Tab. 3. Wyniki analizy zasobów naukowych w wybranych bibliotekach. Stan na marzec 2015 r.

Nazwa biblioteki cyfrowej	**Liczba dostępnych obiektów cyfrowych**	**Liczba gazet i czasopism**[1]	**Liczba obiektów należących do domeny publicznej**[2]	**Liczba domniemanych współczesnych prac naukowych wg analiz szczegółowych**[3]
I	**II**	**III**	**IV**	**V**
Jagiellońska Biblioteka Cyfrowa	270 875	248 973 opisane jako: czasopismo	255 356 opisane jako: domena publiczna (public domain)	Odnaleziono zapis w polu Prawa: copyright 6036; licencja CC 200. To może wskazywać, że są to pozycje naukowe. Zidentyfikowano kolekcję UJ: 4442. Razem 6 036.
Wielkopolska Biblioteka Cyfrowa	251 896	gazeta: 59 559 czasopismo: 48 679 gazety: 45 976 czasopisma: 45 389	jako domena publiczna oznaczono jedynie 25 078 opisów, inne obiekty nie mają w ogóle oznaczeń statusu prawnego.	Zidentyfikowano kolekcję materiały dydaktyczne: 4327. Znaleziono typy: monografia 1575; monografie 631; rozprawa doktorska 1054; artykuł 779; artykuły 391; oprawa habilitacyjna 583; podręczniki akademickie 288; podręczniki 152. Razem 5453.

I	II	III	IV	V
e-biblioteka Uniwersytetu Warszawskiego	204 005	201 367 czasopismo	140 193 domena publiczna	W polu prawa znaleziono zapis: wszystkie prawa zastrzeżone dla 348 pozycji oraz około 150 na innych licencjach, co może wskazywać na prace współczesne i naukowe. Jest wyodrębniona naukowa kolekcja e-publikacje 45. Razem 443.
Śląska Biblioteka Cyfrowa	137 099	czasopismo 108 525	domena publiczna: 60 000 obiektów, nie wszystkie jednak mają opis	Jest kolekcja materiałów dydaktycznych i naukowych 9964. Znaleziono typy: monografia historyczna 190; opracowanie naukowe 181; opracowanie statystyczne 4148; artykuł 1151; rozprawa doktorska 436.
Małopolska Biblioteka Cyfrowa	88 521	czasopismo 78 323	jako domena publiczna oznaczono jedynie 17 817, nie wszystkie obiekty są opisane	Oznaczono typ: artykuł 137. Oznaczono także osobno kolekcję naukową i dydaktyczną 23 691. Zakładamy, że typ artykuł znalazł się w kolekcji naukowej.
Kujawsko-Pomorska Biblioteka Cyfrowa	80 548	czasopismo 68 890	domena publiczna 69 534	Oznaczono także pola prawa z wartościami: wszystkie prawa zastrzeżone 9925; nie wszystkie prawa zastrzeżone 312 czyli licencje CC. Wydzielona kolekcja materiały dydaktyczne 11 038.
Biblioteka Cyfrowa Uniwersytetu Wrocławskiego	52 427	czasopisma 22 614 wydawnictwa ciągłe 8 691	domena publiczna 50 174	Znaleziono około 1000 obiektów, które mają oznaczenia copyright. Jest kolekcja publikacje współczesne 711; e-czytelnia, obiekty współczesne zamknięte 164, materiały edukacyjne 195, e-książki 147. Razem 1217.
Repozytorium Cyfrowe Instytutów Naukowych	52 197	czasopismo 593	domena publiczna 9 350	Odnaleziono opis w polu PRAWA: zastrzeżone – dostęp ograniczony 21 131; prawa zastrzeżone – dostęp nieograniczony 21 087; prac na licencjach CC jest około 300. Zakłada się, że w tym repozytorium wszystkie prace są naukowe. Należy odjąć obiekty z domeny publicznej. Razem 42 847.
Bałtycka Biblioteka Cyfrowa	43 910	czasopismo 25 760 gazeta 2 892	Domena publiczna 5 293 Domena Publiczna 43 684	Odnaleziono kolekcję artykuły 126. W opisach także są określenia: Licencja instytucjonalna 412; licencja instytucjonalna 30 (mała litera). Razem: 126.

I	II	III	IV	V
Mazowiecka Biblioteka Cyfrowa	31 218	czasopism 27 914	domena publiczna 30745	Nie ma wydzielonej kolekcji naukowej, pojedyncze prace naukowe, artykuły: 3 Razem: 3
Zachodniopomorska Biblioteka Cyfrowa „Pomerania”	31 168	czasopismo 21 015	domena publiczna 6 994 spora część zasobów została nieprawidłowo przypisana poszczególnym bibliotekom	Znaleziono typy: dokument naukowo-dydaktyczny 148; praca doktorska 91; materiały konferencyjne 10; prezentacja multimedialna 17. Są kolekcje: nauka, dydaktyka 424; doktoraty i habilitacje 105; mat. konf. 19. Razem: 548.
Dolnośląska Biblioteka Cyfrowa	25 153	czasopismo 9784	brak określenia domena publiczna	Znaleziono typy: rozprawa doktorska 472; doktorat 6; materiały konferencyjne 52, artykuły 3857; rozprawy habilitacyjne 74; recenzje 27; raporty 3; bibliografie 15; książki (wiele współczesnych opracowań) 3986; wykład 1. Razem: 8493
Biblioteka Cyfrowa UMCS	16 889	czasopismo 1129	domena publiczna 10 417	Odnaleziono kolekcje: nauka i dydaktyka 4560; Zidentyfikowano typy: artykuł 1923; rozprawa doktorska 5; monografia 1; praca mgr 1; dyplom 12; referat 4; dokument elektroniczny 7.
Zielonogórska Biblioteka Cyfrowa	14 266	czasopismo 5928 gazeta 1951	brak danych	Istnieje kolekcja nauka i dydaktyka 830; habilitacje i doktoraty 128; Odnaleziono typy: rozprawa doktorska 69; artykuł 65. Razem: 958.
Repozytorium Uniwersytetu im. Adama Mickiewicza	10 741	czasopisma 8601	brak danych	Repozytorium instytucjonalne. Wszystkie prace są naukowe współczesne, jest to dorobek UAM: 10 741 pozycji[5].

[1] W celu wykluczenia ich z dalszych badań, z powodów wyjaśnionych w końcowej części podrozdziału 4.1. Przetwarzanie metadanych – pole TYP.

[2] J.w.

[3] W bardzo dużym zaokrągleniu i przy dużej niepewności.

[4] Literówka (duża litera P) powoduje, że wyszukiwanie trzeba prowadzić dwiema ścieżkami.

[5] Zawartość repozytoriów instytucjonalnych łatwo poddaje się analizie, ponieważ występują tam wyłącznie współczesne prace badawcze.

W analizie eksperckiej przebadano 29 największych bibliotek cyfrowych (w dwóch brak było danych powiązanych z celem badania) oraz repozytoriów naukowych, wśród których wskazano 176 147 domniemanych obiektów naukowych. Uwagi przedstawione w kolumnie *V: liczba domniemanych współczesnych prac naukowych* pokazują, że nawet taka pogłębiona analiza nie daje pewności co do rezultatów poszukiwań, za dużo w nich danych o niepewnych wartościach. Nie ma pewności, czy sumując zasoby nie popełniono błędu rachunkowego, wynikającego z tego, że bibliotekarze przypisali jeden obiekt do dwóch różnych cytowanych kolekcji czy typów. Nie wszystkie określenia typów są jednoznaczne, np. książka czy artykuł, mogą oznaczać prace naukowe, ale mogą też opisywać, np. beletrystykę. Ponadto w kolumnie II, gdzie wyodrębniono czasopisma, wiele z nich należy do domeny publicznej, ale sporo z nich jest naukowych. Zawierają one artykuły, których w metadanych nie można zidentyfikować, w większości nie brano ich zatem pod uwagę. Liczba ta jest tylko szacowana, nie można jej przyjąć za pewnik.

Na podstawie wyników badań eksperckich zaprezentowanej próbki polskich bibliotek cyfrowych można przedstawić następujące wnioski:

(1) Zdecydowana większość zasobów dostępnych za pośrednictwem FBC (ok. 80%) stanowią utwory, które trafiły do domeny publicznej z racji wieku, co niestety nie znajduje odzwierciedlenia w metadanych z powodu występowania nieprawidłowości w wypełnianiu przez poszczególne biblioteki pola *dc:rights*. Wśród zasobów z domeny publicznej ogromną część stanowią czasopisma, skanowane ze względu na stan zachowania i regionalną ważność, co pokazują dość precyzyjnie niektóre dane statystyczne dostarczane przez biblioteki.
(2) Z pozostałych 20% zasobów przeważają te o charakterze nienaukowym (dokumenty życia społecznego, fotografie, grafika, mapy i inne).
(3) Większość spośród prac naukowych stanowią prace spoza zakresu nauk humanistycznych i społecznych (w zasobach politechnik zdecydowaną większość utworów stanowią prace z zakresu nauk technicznych i ścisłych).

Według danych za luty 2016 r. FBC (2016a) udostępniało ok. 1.2 mln obiektów dostępnych bez żadnych ograniczeń, co oznacza, że mogą one należeć do domeny publicznej. Kolejny milion dokumentów miało status prawny albo nieoznaczony, albo było chronione. W całym zasobie znajdowało się ponad 1 740 000 czasopism. Artykułów było nieco ponad 110 000, rozpraw doktorskich 12700, książek 138 000, rozpraw habilitacyjnych 1540, zaś prac dyplomowych 247 (wszystkie dane wyodrębniono na podstawie *dc:type*). Oznacza to, że potencjalnych prac naukowych w FBC było w lutym 2016 r. co najmniej 14 487, zaś maksymalnie mogło ich być około 26 2487.

6. Wnioski

Analiza metadanych udostępnianych przez polskie biblioteki i repozytoria cyfrowe, zarówno automatyczna jak i ekspercka, jest bardzo trudna. Jednym z powodów tej sytuacji jest niezadowalająca dostępność danych. Dla przykładu, największy polski agregator metadanych z bibliotek cyfrowych – FBC udostępnia dane w postaci plików wyeksportowanych z bazy danych, zaś przez protokół OAI-PMH „może udostępniać wybranym podmiotom dane gromadzone z polskich instytucji nauki i kultury" (FBC, 2016b). Utrudnia to automatyczną aktualizację wyników przeprowadzanych badań.

Najważniejszym jednak problemem, jaki spotykają badacze chcący oprzeć swoje badania na metadanych polskich bibliotek cyfrowych, jest niekonsekwencja panująca w opisach bibliograficznych. Większość zasobów polskich bibliotek cyfrowych i repozytoriów opisana jest za pomocą standardu Dublin Core, ale istnieją również biblioteki, które stosują swój własny model i nie zamierzają go zmieniać, ponieważ wymagałoby to wiele pracy, szczególnie przy retrokonwersji. Również w zakresie danych z bibliotek stosujących w opisie dokumentów standard Dublin Core, nie ma zachowanej zgodności we wprowadzaniu danych do poszczególnych pól. W znacznej części dane nie są spójne nie tylko na poziomie pól opisu, ale i zawartości poszczególnych pól.

Pomimo znaczącego przyrostu liczby udostępnianych zasobów cyfrowych od 2009 r. stwierdzono, że w zakresie jakości metadanych nie zaszła znacząca zmiana, w porównaniu do ocen, wniosków i postulatów zaprezentowanych wcześniej przez A. Kazana i E. Skubałę (2008) oraz J. Potęgę (2009).

W niniejszym artykule zaprezentowano wyniki analiz zawartości tylko trzech pól schematu metadanych Dublin Core. Wskazane błędy oraz brak konsekwencji, różne formy językowe wpisów, czy brak kontroli jakości (np. literówki) występują również w pozostałych polach metadanych, co zostało potwierdzone w innej części badań, niezaprezentowanej w niniejszym artykule. Dla przykładu w polu *dc:coverage* spotkać można opis dokumentu, który powinien znajdować się w polu *dc:description* lub słowa kluczowe, które powinny widnieć w polu *dc:keywords*.

W kontekście publikacji M. Werli (2010) warto zasygnalizować, że autorzy niniejszego artykułu, na podstawie analizy metadanych z polskich bibliotek cyfrowych, doszli do podobnych dotyczących jakości opisów w polach *Data*.

W przypadku pola *dc:type* utworzenie wzorcowego zbioru haseł opisujących potencjalne prace naukowe wymagało analizy metadanych z wszystkich dostępnych bibliotek cyfrowych. Uwzględniając skończoną liczbę typów dokumentów, które rzeczywiście opisują prace naukowe, jest to sytuacja kuriozalna, ponieważ w różnych bibliotekach różnie nazywane są takie same typy dokumentów.

Ogromna różnorodność i niekonsekwencje w stosowaniu terminów wpisywanych w pola standardu DublinCore w opisach bibliograficznych z polskich bibliotek cyfrowych wynika, według oceny autorów, z faktu, że bibliotekarze samodzielnie definiują zawartość pól opisu bibliograficznego, wypełniając te pola bez użycia słownictwa kontrolowanego. W Polsce nie zaadaptowano do tej pory słowników kontrolowanych na użytek bibliotek cyfrowych i repozytoriów czy archiwów. Powstałe wiele lat temu Centrum Kompetencji Digitalizacji dla bibliotek, działające w Bibliotece Narodowej, nie potrafi skoordynować działań w zakresie wypracowania standardów, choć należy to do jego obowiązków statutowych i dostaje na ten cel środki finansowe. Warto przypomnieć, że do zadań Centrów Kompetencji należą (MKiDN, 2016):

- *wdrażanie zmian technologicznych dotyczących digitalizacji i przechowywania danych cyfrowych;*
- *koordynacja w zakresie gromadzenia i przechowywania zasobów cyfrowych;*
- *edukacja kadr instytucji kultury prowadzących digitalizację;*
- *udostępnienie materiałów zdigitalizowanych;*
- *wypracowanie standardów;*
- *promocję zasobów cyfrowych.*

Niestety Centrum nie podjęło dotąd żadnych konkretnych inicjatyw w zakresie koordynacji działań bibliotek cyfrowych.

Na niską jakość metadanych polskich bibliotek cyfrowych ma niewątpliwy wpływ brak konsekwencji w stosowaniu obowiązujących w danej bibliotece standardów. Również dopuszczanie do występowania synonimów na listach słownictwa kontrolowanego jest działaniem niedopuszczalnym. Listy takie powinny obowiązywać w każdej bibliotece oraz podawać terminy jednoznaczne, bez synonimów. Z punktu widzenia automatycznego przetwarzania i analizy danych istotne jest, aby dane reprezentujące takie same wartości były w ten sam sposób zapisywane. Można przygotować system automatycznej analizy tak, aby rozpoznawał wszystkie możliwe wariancje typów zapisu danych w poszczególnych polach, ale wymaga to wnikliwej, eksperckiej analizy zbioru metadanych, dodatkowo wykonywanej odrębnie dla każdego osobnego źródła danych, co niepotrzebnie podnosi koszt analiz automatycznych.

Analizy kolejnych pól metadanych obiektów cyfrowych udostępnianych przez polskie biblioteki cyfrowe wiodą do pesymistycznej refleksji, że instytucje odpowiedzialne za budowanie i udostępnianie cyfrowego dziedzictwa wydają się być nieświadome kwestii jakości danych, koncentrując się głównie na istnieniu oraz liczebności tych danych. Niestety, niepoprawne dane, nawet dostępne w największych ilościach, są badawczo nieprzydatne – na ich podstawie nie można metodami analizy automatycznej oraz statystyki wyciągnąć żadnych wiarygodnych, wartościowych wniosków. Oprócz wniosku dotyczącego jakości danych, oczywiście.

Wydaje się, że pracownicy polskich bibliotek cyfrowych nie mają jeszcze świadomości wartości metadanych w erze BigData, a opisy obiektów cyfrowych są przygotowywane nie jako niezależne dokumenty cyfrowe, mające wartość badawczą, ale nadal jako katalogowe reprezentacje dokumentów, jak ma to miejsce w tradycyjnych bibliotekach.

Sytuację ratuje fakt bardzo powszechnego stosowania systemu dLibra[19], który promuje wykorzystanie standardu Dublin Core oraz zapewnia jednolity zapis plików z metadanymi, co znacznie ułatwia ich automatyczne przetwarzanie i analizy.

7. Podsumowanie

Autorzy niniejszego artykułu zdają sobie sprawę z tego, że wnioski wyciągnięte z dotychczasowych badań nie są zadowalające. Poszukują więc nadal metod i narzędzi, które mogą poszerzyć zakres analiz i wnioskowania oraz wnieść nowe rozwiązania. Dotychczasowe analizy pokazują wielorakość problemów wynikającą z jakości pozyskanych danych, problemy te rozwiązywane są przy pomocy metod automatycznych albo za pomocą eksperckich analiz jakościowych. W celu przeprowadzenia automatycznych analiz na dużych zbiorach danych, oprócz stosowania metod autorskich, podjęto współpracę z twórcami projektu CLARIN.PL, m.in. w zakresie automatycznego klasyfikowania dokumentów z zakresu nauk humanistyczno-społecznych na podstawie ich opisów bibliograficznych (CLARIN PL, 2016).

Dotychczasowe analizy metadanych polskich bibliotek cyfrowych wykazały, że twórcom tych bibliotek brakuje jeszcze świadomości wagi poprawnych opisów oraz wymagań dla ich

[19] http://dlibra.psnc.pl/

przydatności. Na jakości danych z bibliotek cyfrowych zaważył również brak wypracowanych metod i standardów opisu dokumentów, powodujący dowolność i niekonsekwencję w opisach nawet w ramach jednej biblioteki. Stan bieżący oraz problemy z automatycznych wskazaniem, na podstawie opisu, dokumentów mogących być pracami naukowymi dał autorom impuls do rozpoczęcia prac nad przygotowaniem podręcznego słownika typów dokumentów na użytek bibliotek cyfrowych ale z uwzględnieniem potrzeb badawczych, m.in. naukometrycznych.

Obiecującym rozwinięciem metody grupowania dokumentów na podstawie ich typów jest powiązanie dokumentu z afiliacją autora, jako wspólna wskazówka, że mamy do czynienia z pracą naukową. Badania będą kontynuowane w kierunku usprawnienia metod identyfikacji dokumentów naukowych, a wśród nich praz z zakresu nauk humanistycznych i społecznych. Prace przygotowawcze w tym kierunku zostały już poczynione. Kolejnym etapem są analizy danych z repozytoriów naukowych, dla których z większą dozą pewności można założyć, że gromadzą wyłącznie prace naukowe.

Opracowania przytoczone w rozdziale 1.2 dotyczyły głównie okresu 1995–2011, podczas gdy niniejszy artykuł prezentuje wyniki badań na materiale współczesnym, publikacjach i ich metadanych z lat 2012–2014. W tym okresie można zaobserwować znaczący przyrost dokumentów cyfrowych dostępnych za pośrednictwem FBC: ok. 396 000 do ok. 1320000. Znamienny jest fakt, że pomimo podnoszenia kwestii wagi poprawności metadanych w dostępie do publikacji cyfrowych nadal trzeba wysuwać te same postulaty.

Źródła finansowania:
Badania przeprowadzono w ramach grantu NCN 2013/11/B/HS2/03048.
Badanie przeprowadzono przy współpracy z realizatorami projektu Polska część infrastruktury naukowej CLARIN ERIC CLARIN PL.

Bibliografia

ANSI/NISO Z39.85–2012 The Dublin Core Metadata Element Set [online] NISO, Baltimore 2013. [05.08.2017], http://www.niso.org/apps/group_public/download.php/10256/Z39–85-2012_dublin_core.pdf

Baca, M., ed. (2016). *Introduction to Metadata*. 3rd ed. [online] Los Angeles: Getty Publications, 2016. [05.08.2017] http://www.getty.edu/publications/intrometadata

Bednarek-Michalska, B. (2014). Prawo autorskie i jego ograniczenia dla polskich bibliotek cyfrowych.. W: A. Puławski (red.) *Znaczenie udostępniania kopii cyfrowych regionalnych zbiorów bibliotecznych w sieci* : materiały z konferencji, Stargard Szczeciński, 5 września 2014 r. Stargard Szczeciński, 5. September 2014, 51–73.

CLARIN PL (2016). *Polska część infrastruktury naukowej CLARIN ERIC*. [online]. CLARIN PL, [05.08.2017], http://clarin-pl.eu/

DCMI Metadata Terms, [online], 2012 [05.08.2017] http://dublincore.org/documents/dcmi-terms/

DCMI Type Vocabulary, [online], 2000, [05.08.2017] http://dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/

Derfert-Wolf, L. Jak posługiwać się biblioteką cyfrową? W: H. Hollender (red). *Cyfrowy świat dokumentu – wydawnictwa, biblioteki, muzea, archiwa*. Warszawa, CPI 2011, 188–237.

Digitalizacja.pl [online] [05.08.2017] http://www.digitalizacja.pl/

Domagalska, A. (2006). Problemy jakości metaopisów w bibliotekach cyfrowych – II Krajowa Konferencja Naukowa Technologie Przetwarzania Danych [online] [05.08.2017], http://www.cs.put.poznan.pl/kkntpd/tpd_pliki/publikacja/pub/55.pdf

FBC (2016a). *Moduł analityczny FBC*. [online]. Federacja Bibliotek Cyfrowych, [05.08.2017], http://fbc.pionier.net.pl/pro/wp-content/plugins/baza-fbc/pivot/

FBC (2016b). *Otwarte dane FBC – API*. [online]. Federacja Bibliotek Cyfrowych, [05.08.2017], http://fbc.pionier.net.pl/pro/wspolpraca/api/

Głowacka, E. (2011). Jakość bibliotek cyfrowych – aspekty i kryteria oceny. e-mentor. Dwumiesięcznik Szkoły Głównej Handlowej w Warszawie [online], 2011, 2 (39), [05.08.2017], http://www.e-mentor.edu.pl/artykul/index/numer/39/id/828

Kazan, A., Skubała, E. (2008). Polskie biblioteki cyfrowe na platformie dLibra – zasób w kontekście tworzenia nowoczesnych kolekcji źródeł informacji dla nauk technicznych, W: H. Ganińska, (red.) *Informacja dla nauki a świat zasobów cyfrowych*, Poznań 2008, 21–33.

Interpretacja schematu Dublin Core wraz z materiałami pomocniczymi dla redaktorów zasobów cyfrowych Biblioteki Cyfrowej Uniwersytetu Wrocławskiego [online] Wrocław 2006 [05.08.2017] http://www.bu.uni.wroc.pl/sites/default/files/images/doc/bc/eporadnik_redaktora_bcuwr.pdf

ISO 8601. Date and Time Formats. [online]. W3C. [05.08.2017] https://www.w3.org/TR/NOTE-datetime

KPBC (2016). *Opis kolekcji: Materiały dydaktyczne*. [online]. Kujawsko-Pomorska Biblioteka Cyfrowa, [05.08.2017], http://kpbc.umk.pl/dlibra/collectiondescription?dirids=1

Metadane, zagadnienia słowników kontrolowanych (Kołpanowicz M., red.) [online] Narodowy Instytut Muzealnictwa i Ochrony Zbiorów, 2012. [05.08.2017] http://nimoz.pl/upload/digitalizacja/Raport_Metadane_NIMOZ_2012.pdf

MKiDN (2016). *Digitalizacja. Działalność Centrów Kompetencji*. [online]. Ministerstwo Kultury i Dziedzictwa Narodowego, [05.08.2017], http://www.digit.mkidn.gov.pl/pages/zasoby/centra-kompetencji.php

Nahotko, M. (2010). *Automatyczne tworzenie metadanych*. Bibliotheca Nostra: śląski kwartalnik naukowy 2/2, 13–31 [online] [05.08.2017], http://bazhum.muzhp.pl/media/files/Bibliotheca_Nostra_slaski_kwartalnik_naukowy/Bibliotheca_Nostra_slaski_kwartalnik_naukowy-r2010-t2-n2/Bibliotheca_Nostra_slaski_kwartalnik_naukowy-r2010-t2-n2-s13–31/Bibliotheca_Nostra_slaski_kwartalnik_naukowy-r2010-t2-n2-s13–31.pdf

Nahotko, M. (2000). *Metadane* W: EBIBB 6/2000(14) [online] [05.08.2017], http://www.oss.wroc.pl/biuletyn/ebib14/nahotko.html

Osińska, V.; Malak, P. (2016a). Dynamiczne sieci społeczne. W: A. Kwiatkowska, M. Sysło. (red.) *Informatyka w edukacji*. Toruń: UMK, 2016.

Osińska, V.; Malak, P. (2016b). Maps and Mapping in Scientometrics. W: M. Góralska, A. Wendel (red.): *Metody i narzędzia badań piśmiennictwa cyfrowego i jego użytkowników*, Wrocław 2016, 59–73.

Osińska, V.; Malak, P.; Bednarek-Michalska, B. (2016). Badanie struktury i dynamiki zasobów cyfrowej wiedzy przy pomocy metod wizualizacji – projekt realizowany na UMK W: R. Bomba, A. Radomski, E. Solska (red.) *Humanistyka Cyfrowa. Badanie tekstów, obrazów i dźwięku*. Lublin e-naukowiec.eu. 8–18, [online] [05.08.2017], http://e-naukowiec.eu/wp-content/uploads/2016/05/Humanistyka_cyfrowa.pdf

Płoszajski, G., red. (2008). Standardy w procesie digitalizacji obiektów dziedzictwa kulturowego, [online] Biblioteka Główna Politechniki Warszawskiej [05.08.2017] http://bcpw.bg.pw.edu.pl/Content/1262/BG_Stand_w_proc_digit.pdf

PN (1992). PN-92-N-01227 — *Bibliotekarstwo i bibliografia. Typologia dokumentów. Terminologia*. Warszawa PKN.

PN-ISO 15836:2006 Informacja i dokumentacja – – Zestaw elementów metadanych Dublin Core (2012). Warszawa PKN.

Potęga, J. (2009). Metadane w polskich bibliotekach cyfrowych W: *Cyfrowość bibliotek i archiwów*. Warszawa, 26–27 listopada 2009 [online]. Biblioteka Narodowa [05.08.2017], http://www.bn.org.pl/download/document/1260454699.ppt

Werla, M. (2010). Wykorzystanie metadanych z polskich bibliotek cyfrowych [online]. W: C. Mazurek, M. Stroiński, J. Węglarz (red.). *Polskie Biblioteki Cyfrowe 2010. Materiały z konferencji zorganizowanej w dniach 20–21 października 2010 roku przez: Bibliotekę Kórnicką PAN, Poznańską Fundację Bibliotek Naukowych, Poznańskie Centrum Superkomputerowo-Sieciowe*. Poznańskie Centrum Superkomputerowo-Sieciowe, Poznań 2011, 125–129 [05.08.2017] http://lib.psnc.pl/Content/376/BC-22-Werla.pdf

Woodward, E. *Metadata for image collection. Inverse proportions: the quantity vs. quality conundrum* [online] American Libraries Magazine – July 21, 2014 [05.08.2017] www.americanlibrariesmagazine.org/article/metadata-image-collections

Analyzing the Metadata Quality in Polish Digital Libraries

Abstract

Purpose/Thesis: The authors discussed the research on the quality of metadata in Polish digital libraries and the evaluation of the possibility to use metadata in automatic processing of digital repositories to detect current scientific documents.
Approach/Methods: The research was conducted in a hybrid way: automatic NLP analysis and expert analysis.
Results and conclusions: The result of the research in question is the identification of the set of failures and misstatements in the metadata of digital objects available in Polish digital libraries.
Practical implications: The improvements to the current metadata state were proposed. The results and conclusions of this research may contribute to the improvement of the quality of metadata and open the possibility of using metadata as valuable research data.
Originality/Value: To the authors' best knowledge, there is not any similar research in Poland as for the scope and scale.

Keywords

Digital libraries. Dublin Core. Metadata analysis. Metadata quality.

dr PIOTR MALAK – adiunkt w Instytucie Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Wrocławskiego. Wcześniej, w latach 2001–2016 pracował jako asystent, a później adiunkt w Instytucie Informacji Naukowej i Bibliologii UMK w Toruniu. Doktor nauk humanistycznych w zakresie bibliologii. Jego zainteresowania badawcze dotyczą inżynierii lingwistycznej, zarządzania informacją, wyszukiwania informacji oraz zarządzania czasem i zadaniami. Stypendysta Szwajcarskiego Funduszu Stypendialnego SCIEX. Członek Polskiego Towarzystwa Informatycznego, International Society for Knowledge Organization, recenzent projektów COST. Autor m.in.: Indeksowanie treści. Porównanie skuteczności metod tradycyjnych i automatycznych, Warszawa: Wydawnictwo SBP 2012; Malak P. Problemy lingwistyczne we współczesnej informacji naukowej: Babik W. (red) Nauka o Informacji. Warszawa: Wydaw. SBP 2016, 469–491; Malak P., Pawłowski A.: Ewaluacja skuteczności systemów wyszukiwania informacji. Od eksperymentu Cranfield do laboratoriów TREC i CLEF. Geneza i metody: Toruńskie Studia Bibliologiczne 2015, 8(2), 137–156.

Kontakt z autorem:
piotr.malak@uwr.edu.pl
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytet Wrocławski

pl. Uniwersytecki 9/13
50–137 Wrocław

dr hab. VESLAVA OSIŃSKA – adiunkt w Instytucie Informacji Naukowej i Bibliologii UMK w Toruniu. Doktor nauk humanistycznych w zakresie bibliologii. Zainteresowania badawcze skupia wokół metod i technik wizualizacji informacji i wizualizacji nauki, a w szczególności analizy dynamiki rozwoju nauki Polskiej. Kierownik grantu badawczego z NCN pt. „Badanie struktury i dynamiki cyfrowych zasobów wiedzy za pomocą metod wizualizacji" (wizualizacjanauki.umk.pl). Członek Management Commettee sieci projektów COST Action TD1201 (knowescape.org). Redaktor portalu www.wizualizacjainformacji.pl. Członek Polskiego Towarzystwa Informatycznego, International Society of Knowledge Organization oraz Stowarzyszenia Naukowców Polaków Litwy. Więcej na stronie: www.umk.pl/~wieo. Autorka: V. Osinska, G. Osinski & B. Kwiatkowska. Visuallization in Learning: Perception, Aesthetics and Pragmatism. In A. Ursyn (Ed.) Maximizinq Coqnitive Learninq throuqh Knowledqe Visualization. Hershey, PA: IGI Global 2015, pp. 381–414; Visual mining czyli eksploracja informacji za pomocą graficznych reprezentacji. Praktyka i Teoria Informacji Naukowo Technicznej 2013, t. 3; Osinska V. Prezentacja informacji: Babik W. (red) Nauka o Informacji, SBP, Warszawa 2016, s. 577–598.

Kontakt z autorką:
wieo@umk.pl
Instytut Informacji Naukowej i Bibliologii
UMK w Toruniu
ul. Bojarskiego 1
87–100 Toruń

mgr BOŻENA BEDNAREK-MICHALSKA – starszy kustosz dyplomowany, zastępca Dyrektora Biblioteki Uniwersyteckiej w Toruniu ds. Informacji i Innowacji. Pracuje jako ekspert doradzający MNiSW w zakresie nowoczesnych technologii informacyjnych, modeli naukowych open access. Była członkini zespołu ds. digitalizacji przy MKiDN oraz zespołu interdyscyplinarnego do spraw działalności upowszechniających naukę przy MNiSW i zespołu ds otwartego dostępu do nauki. Interesuje się głównie tendencjami rozwojowymi w bibliotekach akademickich, nowoczesną komunikacją naukową, bibliotekami cyfrowymi, technologiami informacyjnymi, otwartą nauką i nowymi modelami publikowania naukowego oraz prawem autorskim. Zrealizowała wiele grantów i projektów dla Biblioteki Uniwersyteckiej w Toruniu. Działała w polskich i zagranicznych organizacjach pozarządowych promujących otwartość w nauce (EBIB, KOED, EIFL, E-LIS, SPARC EUROPE) i podnoszenie kwalifikacji bibliotekarzy (Stowarzyszenie EBIB, SBP). Od 1998 r. redaktor naczelna Biuletynu EBIB – czasopisma fachowego open access dla specjalistów informacji i bibliotekarzy, a także redaktorka i wiceprzewodnicząca elektronicznego serwisu EBIB dla bibliotekarzy. Autorka: Ocena jakości bibliotekarskich serwisów informacyjnych udostępnianych w Internecie. Ebib 2, 31, 2002; et. all: Przewodnik po otwartej nauce, ICM UW 2009.

Kontakt z autorką:
Bozena.Bednarek-Michalska@bu.umk.pl
Biblioteka Uniwersytecka, UMK w Toruniu
ul. Gagarina 13
87–100 Toruń