PRZEGLĄD BIBLIOTECZNY 2022 z. 1
PL ISSN 0033-202X

MARCIN ROSZKOWSKI
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski
e-mail: m.roszkowski@uw.edu.pl
ORCID 0000-0001-7396-4685

BIBLIOGRAPHIC DATA SCIENCE – KONCEPTUALIZACJA OBSZARU BADAWCZEGO

Marcin Roszkowski, dr, adiunkt w Katedrze Informatologii Wydziału Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Jego zainteresowania badawcze obejmują problematykę organizacji wiedzy w środowisku sieciowym, w tym metadane oraz ontologie bibliograficzne. Jest autorem kilkudziesięciu publikacji naukowych, w tym m.in. *COVID-19 and the social organization of knowledge in Wikipedia: a study of social representations* „Journal of Documentation", 2021, *The Sociological and Ontological Dimensions of the Knowledge Organization Domain on Google Scholar Citations* „Knowledge Organization", 2020, *Dekonstrukcja artykułu naukowego. Ontologie w publikowaniu semantycznym* „Zagadnienia Informacji Naukowej. Studia Informacyjne", 2019.

SŁOWA KLUCZOWE: Bibliographic data science. Cyfrowa humanistyka. Katalogi biblioteczne. Bibliograficzne bazy danych. Harmonizacja danych.

ABSTRAKT: **Teza/cel artykułu** – Przedmiotem artykułu jest nowy obszar badawczy o nazwie *bibliographic data science*, który charakteryzuje się zastosowaniem metod i technologii *data science* w badaniach nad zasobami katalogów bibliotecznych i bibliografii. Celem artykułu jest próba konceptualizacji *bibliographic data science*. **Metody** – W warstwie metodologicznej przeprowadzone badania opierają się na analizie i krytyce piśmiennictwa dokumentującego badania prowadzone pod szyldem *bibliographic data science* oraz koncepcji analizy domen. **Wyniki** –

U podstaw *bibliographic data science* leży pragmatyczna postawa wobec badań zasobów bibliograficznych z wykorzystaniem metod *data science* realizowanych w humanistyce cyfrowej. W ramach tego obszaru badawczego stawia się problemy właściwe dla dyscyplin tworzących cyfrową humanistykę oraz podejmuje zagadnienia metodologiczne związane z optymalizacją jakości danych bibliograficznych oraz ich harmonizacją. Silne związki *bibliographic data science* z cyfrową humanistyką widoczne są również w społeczności skupionej wokół tego obszaru badawczego.

WSTĘP

Źródła informacji bibliograficznej pełnią kluczową rolę w dokumentacji wytworów dziedzictwa kulturowego oraz wyników działalności badawczej, a także w procesie pośredniczenia między zbiorami informacji a ich użytkownikami. W zależności od celu działalności bibliograficznej mamy do czynienia z różnymi rodzajami kolekcji bibliograficznych. Z jednej strony, są to katalogi biblioteczne (w tym katalogi centralne), których głównym zadaniem jest poinformowanie użytkowników o dokumentach będących w posiadaniu danej instytucji, z drugiej – bibliografie, które rejestrują dokumenty według przyjętych kryteriów (np. formalnych, treściowych) pełniąc funkcje utylitarne, ale również historyczno-archiwizacyjne, prezentując stan szeroko rozumianej kultury w określonym czasie i miejscu (Woźniak-Kasperek, 2015, s. 524). Zasadniczą zawartość źródeł informacji bibliograficznej jako specyficznej formy komunikacji społecznej (Nowak, 2016, s. 6), stanowią dane bibliograficzne, czyli reprezentacje dokumentów tworzone przez wyspecjalizowaną kadrę w oparciu o obowiązujące normy i standardy. Standaryzacja procesu tworzenia metadanych odnosi się zarówno do ich zawartości (zasady katalogowania, kartoteki haseł wzorcowych), jak i do formy prezentacji (format danych bibliograficznych), co w założeniu zapewnia ich wysoką jakość i wiarygodność. Dzięki temu źródła informacji bibliograficznej są wykorzystywane również w działalności badawczej, obecnie często w postaci danych badawczych będących przedmiotem analiz ilościowych.

Zastosowanie metod ilościowych w badaniach zasobów bibliograficznych nie jest jednak czymś nowym. Równolegle do działalności bibliograficznej prowadzono bowiem analizy statystyczne produkcji wydawniczej, które określano mianem statystyki księgoznawczej lub bibliograficznej. Geneza takich badań sięga konwencji berneńskiej (1886) dotyczącej praw autorskich oraz późniejszych prac Międzynarodowego Instytutu Bibliograficznego i w końcu działalności Międzynarodowej Federacji Stowarzyszeń i Instytucji Bibliotekarskich (IFLA) (Sadowska, 2018, s. 192). Od ponad pół wieku prowadzone są również badania bibliometryczne z zastosowaniem metod matematycznych i statystycznych do analizy obiegu

informacji w nauce. Dynamiczny rozwój technologii informacyjno-komunikacyjnych sprawił, że obecnie mamy do dyspozycji szeroką paletę nowych metod i narzędzi wspomagających zarówno proces automatycznego tworzenia metadanych (np. technologie automatycznego indeksowania), jak i ich przetwarzania i analizy, szczególnie w dużej skali (np. przetwarzanie języka naturalnego, uczenie maszynowe, technologie sztucznej inteligencji). To z kolei spowodowało intensyfikację technologicznie wspomaganych badań z wykorzystaniem kolekcji bibliograficznych. Na początku XXI w. pojawił się nowy obszar zastosowania metod ilościowych w badaniach nad zasobami bibliograficznymi, który nazwano *bibliomining* (Nicholson et al., 2003). W tym przypadku mieliśmy do czynienia z wykorzystaniem technik bibliometrycznych w połączeniu z metodami eksploracji danych (ang. *data mining*) w badaniach nad sposobem korzystania przez użytkowników z zasobów katalogów bibliotecznych (Nicholson et al., 2003) i bibliotek cyfrowych (Czapnik, 2016; Nicholson, 2011). Obecnie, w epoce tzw. czwartego paradygmatu w nauce, opartego na intensywnym wykorzystaniu danych cyfrowych w badaniach naukowych (Sosińska-Kalata, 2018, s. 10) źródła bibliograficzne są również traktowane w kategoriach danych masowych, które bada się z wykorzystaniem technologii właściwych dla *data science*. Sam termin *data science* interpretuje się w kategoriach obszaru badawczego, w którym wykorzystuje się metody statystyczne wspomagane zaawansowanymi technologiami informatycznymi. Celem działalności badawczej w ramach *data science* jest ujawnianie właściwości oraz struktury zjawisk przyrodniczych, ludzkich i społecznych z wykorzystaniem danych (Semeler et al., 2019, p. 773). W tym celu korzysta się m.in. z metod i technik eksploracji danych i tekstu, uczenia maszynowego czy technologii właściwych dla big data. Specyfika *data science* polega również na utylitarnym traktowaniu technologii informacyjnych do rozwiązywania problemów badawczych osadzonych w określonych dyscyplinach naukowych czy obszarach badawczych. Mamy więc do czynienia z zastosowaniem data science, np. w nauce o zdrowiu – *health data science* (Hripcsak et al., 2015), nauce o środowisku – *environmental data science* (Gibert et al., 2018), czy w obszarze finansów – *financial data science* (Giudici, 2018).

W 2019 r. w publikacjach fińskich badaczy (Lahti, Marjanen, et al., 2019; Lahti, Vaara, et al., 2019; Tolonen et al., 2019), zrzeszonych w grupie COMHIS – Helsinki Computational History Group[1], pojawiło się wyrażenie *bibliographic data science* (BDS). Zostało ono użyte do nazwania postawy badawczej przyjętej w badaniach nad historią książki z wykorzystaniem metod i technik *data science* w oparciu o metadane pozyskane z wielu źró-

[1] Helsinki Computational History Group: https://www2.helsinki.fi/en/researchgroups/computational-history.

deł bibliograficznych[2]. BDS zostało tam scharakteryzowane jako zastosowanie algorytmicznie wspomaganych procesów harmonizacji i integracji danych pochodzących z różnych źródeł bibliograficznych na potrzeby badań nad produkcją wiedzy (Lahti, Marjanen, et al., 2019, p. 6). Główną przesłanką zastosowania nowych metod przetwarzania danych i związanych z nimi technologii informacyjnych w badaniach nad historią książki był zdaniem twórców tej koncepcji problem braku kompletności i jakości metadanych, szczególnie w bibliografiach historycznych. Natomiast osadzenie BDS w ramach *data science* miało dać nowe rozwiązania metodologiczne i technologiczne pozwalające rozwiązać te problemy oraz dodatkowo usprawnić procesy analizy danych bibliograficznych.

W przypadku BDS mamy więc do czynienia z zastosowaniem zaawansowanych technologii informacyjnych w badaniach nad zasobami bibliograficznymi. Tym samym BDS może być rozumiane jako zastosowanie metod i technik *data science* w badaniach ilościowych nad dużymi i heterogenicznymi zbiorami zasobów bibliograficznych.

W związku z tym wydaje się zasadne postawić pytanie, co nowego proponuje BDS w kontekście ilościowych badań nad zasobami bibliograficznymi oraz jakie problemy badawcze stawia się w tego rodzaju badaniach. Celem artykułu jest próba konceptualizacji BDS jako obszaru badawczego.

METODY

W warstwie metodologicznej przeprowadzone badania opierają się na analizie piśmiennictwa dokumentującego badania prowadzone pod szyldem *bibliographic data science* oraz koncepcji analizy domen (ang. *domain analysis*) jako ramy teoretyczno-metodologicznej, która wyznaczyła określony sposób interpretacji BDS. Analiza piśmiennictwa została przeprowadzona na podstawie rezultatów wyszukiwania w bazie Google Scholar oraz analizy cytowań do publikacji prezentujących założenia BDS. Wybór tego źródła był podyktowany faktem, że termin *bibliographic data science* jest obecny od niedawna w dyskursie naukowym, a baza Google Scholar daje szanse na większą kompletność wyszukiwania niż, np. Scopus czy Web of Science m.in. poprzez indeksowanie preprintów. Analiza piśmiennictwa miała przede wszystkim charakter eksploracyjny. Jej głównym celem było poznanie sposobów interpretacji BDS oraz zgromadzenie materiału na potrzeby analizy domen.

[2] Wcześniejsze użycie tej nazwy można jednak odnaleźć w serwisie społecznościowym Twitter. W 2012 r. Lorcan Dempsey (związany z OCLC) opublikował wpis (Dempsey, 2012) w którym użył wyrażenia *bibliographic data science* charakteryzując ofertę pracy w firmie Mendeley, odpowiedzialnej za menedżer bibliografii o tej samej nazwie, gdzie zakres obowiązków obejmował m.in. zarządzanie i przetwarzanie społecznie tworzonymi metadanymi.

Przez domenę rozumie się pewien zasób wiedzy współdzielony przez daną społeczność, która podziela określone założenia ontologiczne i epistemologiczne (Hjørland, 2017, p. 441). Domeny wiedzy mogą być interpretowane nie tylko jako dyscypliny naukowe, ale również jako obszary badawcze czy specjalności, których badanie zmierza do zrozumienia sposobów tworzenia nowej wiedzy oraz procesów komunikacji naukowej w ramach społeczności z nimi związanych. W analizie domen wyróżnia się zatem trzy wymiary – ontologiczny, epistemologiczny oraz socjologiczny (Hjørland, 2017; Hjørland & Hartel, 2003). Analiza domen w wymiarze ontologicznym zmierza do ujawnienia teorii oraz pojęć będących przedmiotem działalności badawczej. Wymiar epistemologiczny domeny jest związany ze sposobami pozyskiwania nowej wiedzy i jej charakterem, co wiąże się również z obecnością paradygmatów badawczych oraz metodologicznymi aspektami interpretacji rzeczywistości. Wymiar socjologiczny odnosi się do aktorów i grup społecznych zaangażowanych w domenę (Bawden & Robinson, 2015, p. 93).

KONTEKST BDS

Genezy BDS należy szukać w cyfrowej humanistyce. To właśnie tutaj możemy zaobserwować zainteresowanie nie tylko komputerowo wspomaganą interpretacją tekstów, ale również zastosowaniem nowoczesnych technologii informacyjnych w badaniach nad zasobami katalogów bibliotecznych i bibliografii. Jest to szczególnie widoczne w badaniach z obszaru historii, w tym historii książki (np. Lahti et al., 2015, 2020) i literaturoznawstwa (Pawłowski, Herden, et al., 2021; Underwood, 2020). W tych pracach pytania badawcze są zdeterminowane dziedzinowo, ale przyjmowana postawa poznawcza zakłada możliwość opisu rzeczywistości na podstawie metadanych publikacji. Innymi słowy, przyjmuje się założenie, że badanie artefaktów reprezentacyjnych, jakimi są rekordy bibliograficzne, pozwala tworzyć nową wiedzę na temat rzeczywistości. Kolejna cecha badań prowadzonych w tym nurcie, to duża skala danych badawczych i wielość źródeł ich pozyskiwania. Wiąże się to z chęcią pokazania prawidłowości określonych zjawisk, ale również powoduje określone problemy z jakością i spójnością metadanych będących przedmiotem analiz. I to właśnie ten aspekt jest tutaj szczególnie eksponowany. BDS ma w założeniu wypracować rozwiązania metodyczne w zakresie optymalizacji jakości danych bibliograficznych, ich wzbogacania i analizy na potrzeby badań w obszarze cyfrowej humanistyki. Tak więc u podstaw BDS leży pragmatyczna postawa wobec technologicznie wspomaganych badań z wykorzystaniem zasobów bibliograficznych realizowanych pod szyldem cyfrowej humanistyki. Oprócz ewidentnych związków BDS z ilościowym badaniem obiegu informacji w nauce można tutaj zaobserwować

również pewne relacje z wykorzystywaną w literaturoznawstwie koncepcją *distant reading* autorstwa Franco Morettiego (Moretti, 2013). Idea *distant reading* polega na przyjęciu postawy makro w badaniach tekstów literackich poprzez zastosowanie metod i narzędzi komputerowych do analizy ich zawartości oraz wizualizacji rezultatów. Zakłada się bowiem, że komputerowa analiza tekstów literackich jest w stanie wnieść nowe spojrzenie w badaniach z obszaru historii literatury i krytyki literackiej. Koncepcja Morettiego „czytania na dystans" jest w opozycji do praktyki uważnego czytania (ang. *close reading*) poprzez zastosowanie metod ilościowych w badaniach dużych korpusów tekstów literackich i na tej podstawie uprawomocniania uzyskanych rezultatów. Jest charakteryzowana jako „analiza literatury bez zaglądania do treści poszczególnych dzieł" (Eder, 2014, s. 91). W przypadku BDS mamy do czynienia z jeszcze szerszą perspektywą, bowiem przedmiotem badań są metadane, czyli reprezentacje dokumentów (ich formy i treści), a nie bezpośrednia ich zawartość. Sam Moretti prowadził również tego typu badania z wykorzystaniem metadanych dokumentów, gdzie w oparciu o zasoby bibliografii analizował strukturę tytułów XVIII- i XIX-wiecznych powieści angielskich oraz za ich pomocą identyfikował szczegółowe gatunki literackie (Moretti, 2013).

KONCEPCJA BDS

Założenia BDS zostały przedstawione w artykule pt. *Bibliographic data science and the history of the book (c. 1500-1800)* (Lahti, Marjanen, et al., 2019), który został opublikowany w czasopiśmie z obszaru nauki o informacji i bibliotekoznawstwa – *Cataloging and Classification Quarterly*. BDS jest tam określona jako pewnego rodzaju postawa badawcza (ang. *approach*), w której zasoby bibliograficzne są wykorzystywane nie tylko do celów wyszukiwania informacji, ale przede wszystkim jako dane w badaniach ilościowych nad produkcją wydawniczą. U podstaw BDS leżą następujące założenia:

1. Przedmiotem badań BDS są duże i heterogeniczne zasoby bibliograficzne.
2. Zasoby bibliograficzne (metadane) są traktowane jako obiekty badawcze.
3. Celem BDS jest wypracowanie metod i narzędzi optymalizacji jakości i kompletności metadanych.
4. BDS jest oparte na ilościowych badaniach informacji (metody statystyczne i probabilistyczne).
5. BDS jest oparte na modelu iteracyjnym.

W ramach BDS prowadzi się badania na danych masowych, co jest specyficzne dla data science. Tego rodzaju badania mają na celu zgromadzenie dużych ilości danych, często pochodzących z wielu źródeł. Celem jest

uzyskanie kompletnego obrazu zjawiska reprezentowanego przez zbiory publikacji. We wspomnianym wcześniej artykule (Lahti et al., 2019), zaprezentowano wyniki badań nad ewolucją formatów książek wydawanych w Europie między XVI a XIX w., w których zgromadzono ponad 6 milionów rekordów bibliograficznych pochodzących z czterech bibliografii. Adam Pawłowski wraz z zespołem (Pawłowski, Herden, et al., 2021) przeprowadzili badania nad automatyczną identyfikacją gatunków literackich na podstawie tytułów wydawnictw, w których wykorzystano ponad 1,8 miliona rekordów bibliograficznych pochodzących z bibliografii Biblioteki Narodowej (Przewodnik Bibliograficzny).

W BDS każdy rekord bibliograficzny jest traktowany jako obiekt badawczy (ang. *research object*). Z metodologicznego punktu widzenia oznacza to, że jego indywidualne cechy są podstawą do identyfikacji zmiennych w badaniach ilościowych. To na podstawie elementów strukturalnych rekordu metadanych, czyli wybranych cech publikacji, a także relacji między nimi, przeprowadza się analizy ilościowe. Celem tych analiz jest m.in. zobrazowanie skali lub zmian w czasie określonego zjawiska, które są reprezentowane przez własności publikacji. Uzyskane wyniki są przedmiotem krytycznej analizy opartej na wiedzy dziedzinowej i kontekście historycznym zapewniającym właściwe ramy interpretacyjne.

Istotnym problemem, który pojawia się w tego typu badaniach, jest kwestia jakości i kompletność metadanych pozyskiwanych z wielu źródeł bibliograficznych. Mowa tutaj o różnych formatach danych, konwencjach zapisu informacji w rekordach, które wynikają ze stosowanych zasad katalogowania, ale również błędnych lub brakujących danych. Problemy związane z jakością metadanych wynikają nie tylko z faktu agregowania danych z wielu źródeł, ale również z poziomu formalnej gotowości tych zasobów do realizacji założonych celów badawczych. Semantyka schematów metadanych stosowanych w katalogach bibliotecznych i bibliografiach jest zdeterminowana przez dwie główne funkcje, jakie mają realizować te narzędzia – funkcja metainformacyjna i wyszukiwawcza. Pierwsza zakłada odwzorowanie istotnych cech opisywanego zasobu w celu ich identyfikacji i rozróżnienia, druga wykorzystanie na potrzeby wyszukiwania informacji. Potraktowanie rekordu bibliograficznego jako obiektu badawczego zmienia perspektywę oraz wprowadza nowe oczekiwania wobec jego potencjału informacyjnego. Uprawianie BDS oznacza bowiem wykorzystanie zasobów bibliograficznych poza głównym kontekstem ich funkcjonowania (katalog biblioteczny, bibliograficzna baza danych) i ingerencję w ich strukturę oraz zawartość na potrzeby prowadzonych badań. Ingerencja, o której mowa, jest oparta na przesłankach pragmatycznych, zdeterminowanych założonymi celami badawczymi i polega na wykorzystaniu technik i narzędzi data science do optymalizacji jakości rekordów metadanych. Głównym celem BDS w tym obszarze jest wypracowanie od-

twarzalnych procesów harmonizacji (ang. *reproducible data harmonization*) i analizy danych oraz automatycznych lub półautomatycznych technik przetwarzania i zarządzania danymi (ang. *data curation*).

Harmonizacja danych polega na ujednoliceniu ich formy oraz zawartości zgodnie z przyjętymi konwencjami umożliwiającymi późniejsze ich przetwarzanie. Ma ona zapewnić ich formalną normalizację, która pozwoli z kolei na przeprowadzenie analiz statystycznych. Polega to z jednej strony, na przyjęciu określonego formatu danych i zastosowaniu go dla całego zbioru pozyskanych metadanych, z drugiej – na znormalizowaniu wybranych wartości danych istotnych z punktu widzenia prowadzonych badań. W pierwszym przypadku, oznacza to konieczność ponownego modelowania danych, wyboru adekwatnego formatu danych i przeprowadzenia procesu mapowania danych do nowej postaci. W drugim – mowa jest o ingerencji w wartości metadanych poprzez automatyczne usuwanie błędnych zapisów, normalizację form językowych, a także ekstrakcję dodatkowych informacji z wybranych pól rekordów, np. objętość, format książki. Istotnym elementem tego procesu jest również wykorzystanie zewnętrznych źródeł do automatycznego wzbogacania metadanych (ang. *metadata enrichment*) o nowe informacje lub do weryfikacji ich poprawności. Dotyczy to np. identyfikacji płci dla autorów na podstawie form nazw osobowych, pozyskiwania dodatkowych informacji geolokalizacyjnych dla nazw miejscowych.

Harmonizacja danych jest to więc wieloetapowy proces, na który składa się m.in. analiza pozyskanego zbioru danych, ujednolicanie wartości dla poszczególnych elementów metadanych (usuwanie znaków specjalnych, błędów językowych, rozwiązanie problemu form synonimicznych), identyfikacja i usuwanie zduplikowanych rekordów, uzupełnianie brakujących danych poprzez ekstrakcję informacji z innych pól rekordu lub korzystając z zewnętrznych źródeł, wzbogacanie metadanych (Marjanen et al., 2019, p. 60).

Odtwarzalność tych procedur polega na możliwości ich ponownego wykorzystania przez innych badaczy i ma być zapewniona przez ich publicznie dostępną dokumentację wraz z dostępem do technologii informacyjnych, które tam wykorzystano. W tym celu korzysta się z otwartych repozytoriów danych i kodów źródłowych opartych, np. o system Git i udostępnianych przez wiele platform internetowych (np. GitHub czy GitLab).

W procesach przetwarzania i zarządzania metadanymi wykorzystuje się metody i narzędzia właściwe dla *data science*. Należą do nich m.in. przetwarzanie języka naturalnego w oparciu o metody statystyczne i probabilistyczne, zastosowanie metod i technik automatycznej eksploracji danych (ang. *data mining*) i tekstu (ang. *text mining*), wykorzystanie uczenia maszynowego (ang. *machine learning*) do automatycznej klasyfikacji ele-

mentów danych bibliograficznych, jak również wykorzystanie baz danych umożliwiających odzwierciedlenie grafowej/sieciowej natury powiązań między dokumentami. W tym przypadku nie chodzi tylko o rozwiązanie problemu badawczego poprzez zastosowanie nowoczesnych technologii informacyjnych, ale również o wypracowanie spójnej metodyki pracy nad metadanymi (ang. *metadata curation workflow*) i możliwości jej zastosowania przez innych badaczy poprzez zapewnienie jej publicznej dostępności zarówno w postaci dokumentacji, jak i pakietu narzędziowego[3]. Specyfiką procesów harmonizacji i przetwarzania danych bibliograficznych realizowanych w ramach BDS jest ich iteracyjny charakter. Polega to na wielokrotnym powtarzaniu określonych procedur (iteracje) przez algorytmy przetwarzające dane, aż do osiągnięcia satysfakcjonującego poziomu efektywności. Innymi słowy, nie poszukuje się jednego gotowego rozwiązania wszystkich problemów związanych z jakością metadanych, lecz w kolejnych iteracjach optymalizuje się zastosowane dotychczas rozwiązania dla danego elementu metadanych i przechodzi się do następnego. Chociaż Muriel Foulonneau i Jenn Riley (2008, p. 57) twierdzą, że iteracyjne podejście w kontroli jakości metadanych nie ma końca, to jest to jednak ekonomiczne rozwiązanie w kontekście zakładanych celów badawczych, dla których te procedury są stosowane.

Przedstawiona wyżej specyfika operacji dokonywanych na metadanych w ramach BDS powoduje, że mamy do czynienia ze zjawiskiem zmiany przeznaczenia metadanych (ang. *metadata repurposing*). Polega ono na adaptacji i wykorzystaniu metadanych do celów innych niż te, dla których je opracowano oraz poza kontekstem ich zakładanego funkcjonowania (Deng, 2010; Foulonneau & Cole, 2005). Procesy harmonizacji, optymalizacji jakości oraz wzbogacania metadanych są zdeterminowane przez nowe cele, do których są one wykorzystywane – cele badawcze. Tak jak w przypadku bibliometrii zmiana przeznaczenia metadanych jest związana z szeroko rozumianym badaniem obiegu informacji w nauce, tak w przypadku BDS nowe cele są związane z problemami osadzonymi w polach badawczych historii i literaturoznawstwa eksplorowanych w ramach cyfrowej humanistyki.

WYMIAR ONTOLOGICZNY BDS

W tak przedstawionej koncepcji BDS wymiar ontologiczny można interpretować w kategoriach ujęć teoretycznych oraz stawianych problemów badawczych. Jednak wyizolowanie warstwy ontologicznej szczególnie od epistemologicznej jest trudne, ponieważ istnieją między nimi

[3] Np. pakiet Bibliographica (https://github.com/COMHIS/bibliographica) wykorzystany w badaniach (Lahti, Marjanen, et al., 2019)

złożone relacje (Hjørland 2017, p. 440) i owe trzy wymiary wchodzą z sobą w interakcje o charakterze intelektualnym i społecznym (Bawden & Robinson, 2015, p. 93).

Analiza problemów badawczych podejmowanych w pracach odwołujących się do BDS wskazuje na niejednorodność wymiaru ontologicznego. Jest to spowodowane faktem, że BDS jest osadzona w cyfrowej humanistyce, która odwołuje się do teorii i koncepcji właściwych dla historii, literaturoznawstwa, czy językoznawstwa, które w badaniach szczegółowych są konfrontowane z możliwościami zastosowania nowoczesnych technologii informacyjno-komunikacyjnych. Mamy więc do czynienia z problemami badawczymi i w konsekwencji założeniami ontologicznymi właściwymi dla tych dyscyplin. Dalej przedstawiono kilka przykładów badań ilustrujących tę tezę.

W cytowanym wcześniej artykule pt. *Bibliographic data science and the history of the book (c. 1500–1800)* (Lahti, Marjanen, et al., 2019) prezentującym założenia BDS, autorzy wykorzystali zasoby czterech bibliografii w badaniach nad produkcją wydawniczą w Europie od XVI do XIX w. Dane pozyskano ze szwedzkiej i fińskiej bibliografii narodowej, katalogu centralnego English Short-Title Catalogue oraz bazy Heritage of the Printed Book (Baza Dziedzictwa Książki Drukowanej). Badania dotyczyły dwóch problemów – jak zmieniał się format książki drukowanej na przestrzeni wieków oraz jak przebiegał proces upiśmiennienia (wernakularyzacji) w Europie, ze szczególnym uwzględnieniem języka szwedzkiego i fińskiego.

Celem badań Leo Lahti,Eetu Mäkelä, i Mikko Tolonena (Lahti et al., 2020), była reprezentatywność pełnotekstowej kolekcji XVIII-wiecznych książek wydawanych w Wielkiej Brytanii (Eighteenth Century Collections Online; ECCO) na podstawie zasobów bibliografii The English Short-Title Catalogue (ESTC), która stanowiła źródło metadanych dla zdigitalizowanych pozycji. Przedmiotem badań były metadane pochodzące z obydwu kolekcji, a celem badań było wykrycie poziomu dostępności pełnych tekstów w ECCO zarejestrowanych w bibliografii ESTC. Autorzy przyjęli hipotezy, że kolekcja ECCO może marginalizować określone wydawnictwa i założyli, że może mieć na to wpływ płeć autora, forma wydawnictwa oraz data wydania. Badania przeprowadzono na 227 tys. rekordów z bazy ESTC. Po ekstrakcji nazw osobowych z oznaczenia odpowiedzialności wzbogacono te dane o informacje na temat płci autora oraz dat życia. Dane te pozyskano automatycznie z bazy Virtual International Authority File (VIAF). Następnie, na podstawie opracowanego modelu regresji, który miał za zadanie odwzorować prawdopodobieństwo wystąpienia danego tytułu z ESTC w ECCO, zastosowano technologie programowania probabilistycznego (skrypty w języku Python oraz R) i przeprowadzono proces analizy danych. Cały proces oraz wykorzystana

technologia została udokumentowana i udostępniona publicznie[4], a wyniki przeprowadzonych analiz wskazały na istotny udział tych zmiennych w profilu kolekcji ECCO.

Adam Pawłowski wraz z zespołem przeprowadził szereg badań nad zasobami polskiej bibliografii narodowej („Przewodnik Bibliograficzny") z wykorzystaniem metod lingwistycznych wspomaganych przez technologie eksploracji tekstu (ang. *text mining*), które zrealizowano w ramach grantu Narodowego Centrum Nauki pt. „Metody i narzędzia lingwistyki korpusowej w badaniach bibliografii polskich wydawnictw zwartych z lat 1997-2017". W oparciu o metadane na temat 553 tys. książek wydanych w latach 1997-2017, które pozyskano z bazy Biblioteki Narodowej za pośrednictwem interfejsu programistycznego (API), przeprowadzili badania korpusowe nad tytułami publikacji (Pawłowski, Topolski, et al., 2021). Badania przeprowadzono z wykorzystaniem analizy stylostatystycznej porównując strukturę korpusu mikrotekstów, który opracowano na podstawie tytułów publikacji z zasobem Narodowego Korpusu Języka Polskiego. W badaniach wykorzystano techniki przetwarzania języka naturalnego, które zastosowano do analizy metadanych odpowiedzialnych za rejestrację tytułów publikacji. Uzyskane wyniki pokazały różnice między tymi korpusami oraz dały obraz konstrukcji językowych w tytułach współczesnych publikacji książkowych. Przedmiotem kolejnych badań A. Pawłowskiego wraz z zespołem (Pawłowski, Herden, et al., 2021) był zbiór ponad 1,8 mln rekordów publikacji również pochodzących z zasobów „Przewodnika Bibliograficznego". Celem tych badań była możliwość zastosowania technik uczenia maszynowego na potrzeby automatycznej identyfikacji płci autorów oraz gatunków tekstów (m.in. naukowych i literackich). W tym przypadku również zastosowano metody językoznawcze wspomagane technologiami informacyjnymi w odniesieniu do zawartości rekordów bibliograficznych.

Druga kategoria problemów badawczych podejmowana w ramach BDS bezpośrednio wiąże się z koncepcją metadanych. Tutaj podstawowe założenia ontologiczne leżące u podstaw badań, choć nie zawsze ujawniane, odnoszą się do teorii organizacji wiedzy, w której metadane są traktowane jako artefakty reprezentacyjne (Ceusters, 2012). W takiej interpretacji są to wytwory działalności ludzkiej, których głównym celem jest reprezentacja obiektów rzeczywistości, tutaj w postaci dokumentów. Metadane dokumentu mają więc postać jego charakterystyki wyszukiwawczej, którą tworzą dane reprezentujące sposoby jego utrwalenia (cechy formalne) i/lub treści (Sosińska-Kalata & Roszkowski, 2016, s. 299). W kontekście BDS mamy więc do czynienia z problematyką jakości metadanych rozumianą w kategoriach poprawności, spójności, wiarygodności, kompletności oraz

[4] 2020 Chr Project: https://gitlab.com/COMHIS/2020-chr.

metodami i narzędziami ich optymalizacji. Dalej przedstawiono kilka przykładów z tego nurtu badań.

Evan Bryer wraz z zespołem (Bryer et al., 2021) opracowali metodę deduplikacji rekordów metadanych w oparciu o strategię oczyszczania i analizę skupień z wykorzystaniem metod i technologii uczenia maszynowego. Przedmiotem badań była baza ponad 5 mln rekordów w formacie MARC 21 dla wydawnictw zwartych opublikowanych między XVI a XIX w. i dostępnych za pośrednictwem bazy WorldCat. Identyfikację duplikatów przeprowadzono na podstawie analizy tytułów publikacji. Celem badań było wypracowanie metod optymalizacji jakości metadanych na potrzeby badań nad produkcją wydawniczą.

Andreas Luschow i José Calvo Tello (2021) opracowali metodę identyfikacji wybranych gatunków literackich w katalogach bibliotecznych. Przedmiotem ich badań były zasoby bibliograficzne niemieckiego katalogu rozproszonego Gemeinsamer Verbundkatalog (GBV). Również i tutaj wykorzystano metody i technologie uczenia maszynowego. Punktem wyjścia był wykaz gatunków literackich pobrany z kartoteki Gemeinsame Normdatei. Były one podstawą do konstrukcji modelu, który zastosowano w procesie analizy danych. Zdaniem autorów ich eksperyment badawczy miał na celu wypracowanie najbardziej efektywnego rozwiązania, które można zastosować w badaniach z wykorzystaniem dużych zasobów bibliograficznych.

Bardziej kompleksowe podejście do metod przetwarzania danych bibliograficznych na potrzeby badań naukowych zaproponowali Róbert Péter wraz z zespołem (Péter et al., 2020). Efektem pracy tego zespołu badawczego jest AVOBMAT[5] – platforma do wizualizacji i analizy zarówno zasobów bibliograficznych, jak i pełnych tekstów dokumentów. Projekt jest rozwijany od 2017 r., a samo narzędzie opracowano z wykorzystaniem technologii przetwarzania języka naturalnego oraz sztucznej inteligencji. AVOBMAT pozwala zaimportować metadane z menedżera bibliograficznego Zotero (format CSV lub RDF) oraz zastosować zaawansowane techniki wizualizacji oraz analizy dystrybucji wartości dla wybranych elementów metadanych. Zastosowane metody i narzędzia przetwarzania języka naturalnego uwzględniają obsługę tekstów w 52 językach, w tym w języku polskim. Platforma powstała na potrzeby prowadzenia badań z wykorzystaniem zasobów bibliograficznych.

WYMIAR EPISTEMOLOGICZNY BDS

Dotychczasowe rozważania na temat BDS sytuują je w ramach cyfrowej humanistyki jako technologicznie zdeterminowaną postawę badaw-

[5] AVOBMAT: https://avobmat.hu/.

czą wobec zasobów bibliograficznych, która w warstwie metodologicznej i narzędziowej czerpie z dokonań *data science*. Skoro BDS jest osadzona w cyfrowej humanistyce, to reprezentuje stanowisko epistemiczne właściwe dla tego obszaru.

Pojęcie stanowiska epistemicznego należy rozumieć jako określoną postawę poznawczą wobec procesu pozyskiwania wiedzy (Gkeredakis et al., 2016). Hope Olson (1996) definiuje ten termin w kategoriach poglądu lub przekonania odnoszącego się do natury wiedzy i sposobu jej tworzenia. Tego rodzaju postawa wobec procesu poznania naukowego zakłada określone sposoby pozyskiwania informacji na temat rzeczywistości a także tworzenia i uwiarygodniania nowej wiedzy. Stanowisko epistemiczne jest więc pewnego rodzaju strategią lub połączeniem strategii służących do tworzenia uzasadnionych przekonań (Chakravartty, 2004, p. 175). Przyjęcie określonego stanowiska epistemicznego w badaniach naukowych wpływa bezpośrednio na decyzje o charakterze metodologicznym.

Marija Dalbello (2011) twierdzi, że w cyfrowej humanistyce przyjmuje się stanowisko epistemiczne zakładające metodę rozumowania w oparciu o technologiczne artefakty (p. 481) oraz koncentruje się uwagę na procesie interpretacji z wykorzystaniem cyfrowych narzędzi jako głównych źródeł poznania (p. 497). Hermeneutyczne tradycje humanistyki przyjmują tutaj nową odsłonę w postaci komputerowo wspomaganej interpretacji tekstów, czy nawet korpusów tekstów, co w szerszym ujęciu niewątpliwie nawiązuje do koncepcji cyfrowej hermeneutyki Rafaela Capurro (Capurro, 2010). W przypadku BDS poznanie następuje za pośrednictwem technologii a głównym źródłem poznania są metadane. Te z kolei mają status artefaktów reprezentacyjnych. Tak jak mapa jest artefaktem reprezentacyjnym fragmentu rzeczywistości, tak metadane pełnią taką funkcję wobec dokumentu, którego formę i zawartość opisują. Sprawę można skomplikować zadając pytanie, jaką rzeczywistość reprezentują metadane. W myśl teorii języków informacyjno-wyszukiwawczych, do których zalicza się język opisu bibliograficznego (Sosińska-Kalata & Roszkowski, 2016), desygnatami bezpośrednimi jednostek leksykalnych (elementy metadanych) są dokumenty, a pośrednimi obiekty rzeczywistości pozadokumentacyjnej, których dotyczą dokumenty (Bojar, 2002, s. 48). W perspektywie epistemicznej oznacza to, że BDS proponuje poznanie rzeczywistości za pośrednictwem „podwójnego sita". Bezpośrednim przedmiotem poznania są metadane dokumentu, a ten z kolei pozostaje w jakiejś (tutaj nieokreślonej) relacji do rzeczywistości. Oczywiście gatunek dokumentu (dzieło literackie, praca naukowa) będzie miał tutaj kluczowe znaczenie przy określaniu tego stosunku, ale faktem pozostaje bezpośrednia i pośrednia koncepcja znaczenia przy traktowaniu metadanych jako obiektów badawczych.

Z drugiej strony można odwołać się do koncepcji metadanych jako dowodu lub świadectwa (ang. *evidence*). Matthew S. Mayernik (2019) proponuje taką koncepcję, przyjmując definicję danych jako jednostek pełniących funkcję świadectwa lub dowodu zjawisk na potrzeby działalności badawczej. Dane pełnią funkcję społeczną i retoryczną wobec zjawisk, które reprezentują, a sama reprezentacja jest intencjonalna i z natury selektywna. Według niego metadane są podkategorią danych, a określenie czegoś mianem metadanych ma charakter sytuacyjny, społeczny, a nawet polityczny (s. 733). Wykorzystanie metadanych jako dowodu lub świadectwa oznacza konieczność wprowadzenia pojęcia odpowiedzialności (ang. *accountability*) w ich interpretacji jako procesu i produktu. Oznacza to, że traktowanie metadanych jako danych badawczych wymaga zrozumienia celów, dla których powstały, otoczenia społecznego oraz sposobu ekspresji metadanych z wykorzystaniem standardów i konwencji stosowanych przez ich twórców. W dużym uproszczeniu, chodzi o uwzględnienie zmiennych kontekstowych determinujących formę i zawartość rekordów metadanych. W przypadku BDS mamy do czynienia z metadanymi zastanymi i przejętymi z docelowego kontekstu ich funkcjonowania. Proces poznawania rzeczywistości za ich pośrednictwem wymaga więc zrozumienia nie tylko zawartości zasobów bibliograficznych, ale również ich historycznych i społecznych uwarunkowań, co oznacza szczególną uwagę w stawianiu pytań badawczych oraz uprawomocnianiu uzyskanych rezultatów.

WYMIAR SOCJOLOGICZNY BDS

Wymiar socjologiczny domeny można interpretować w kategoriach zagadnień odnoszących się do grup osób z nią związanych. Birger Hjørland i Hanne Albrechtsen (1995, p. 400) twierdzą, że ten wymiar ma kluczowe znaczenie dla właściwego poznania domeny, ponieważ pozwala spojrzeć na proces tworzenia i komunikowania wiedzy przez pryzmat wspólnoty dyskursywnej (ang. *discourse community*), czyli społeczności, w której zachodzą ustrukturyzowane procesy komunikacyjne i aktywna wymiana informacji w ramach np. współdzielonych celów i poglądów (Hjørland, 2002, p. 258). W przypadku BDS wymiar ten odnosi się do uczestników komunikacji naukowej. Z metodologicznego punktu widzenia wymiar socjologiczny danej domeny można zrekonstruować np. poprzez identyfikację członków danej wspólnoty dyskursywnej z wykorzystaniem technik bibliometrycznych. W przypadku BDS problemem może być jednak metodyka tworzenia korpusu publikacji właściwych dla tej domeny. Z jednej strony, mamy bowiem teksty, w których autorzy wprost odwołują się do nazwy *bibliographic data science* jako przyjętej postawy badawczej, z drugiej – nie można zapomnieć o fakcie, że badania nad zasobami bi-

bliograficznymi z wykorzystaniem technologii *data science* są prowadzone przynajmniej od kilku lat. Na przykład baza Scopus dla zapytania „bibliographic" AND „data mining" oferuje ponad 500 rezultatów wyszukiwania, dla „bibliographic" AND „machine learning" ponad 300 natomiast dla frazy „bibliographic data science" tylko siedem. Z uwagi na rozpoznawczy charakter badań przedstawionych w tym artykule w dalszej części wymiar socjologiczny będzie dotyczył wyłącznie tych badaczy, którzy wprost odwołują się do koncepcji BDS w swoich pracach. Badania eksploracyjne przeprowadzono z wykorzystaniem bazy Google Scholar, co zapewniło większą kompletność rezultatów. Na początku 2022 r. Google Scholar rejestrował 82 pozycje dla frazy „bibliographic data science". Po przeanalizowaniu wszystkich rezultatów do dalszych analiz włączono 25 publikacji. Pozostałe albo były zduplikowanymi zapisami, albo w ich treści występowały tylko wzmianki o BDS jako nowym podejściu w badaniach nad zasobami bibliograficznymi.

W obecnej fazie rozwoju BDS, czyli na początkowym etapie, biorąc pod uwagę fakt, że nazwa ta pojawiła się w 2019 r., społeczność skupiona wokół BDS jest niewielka i tworzy ją grupa 48 badaczy. Są to przede wszystkim przedstawiciele jednostek uniwersyteckich (wydziały, instytuty) lub centrów badawczych związanych bezpośrednio z cyfrową humanistyką (np. Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki; Centre for Contemporary and Digital History (C2DH), University of Luxembourg) lub poszczególnymi dyscyplinami humanistycznymi (np. Institute of English and American Studies, University of Szeged; Faculty of Philosophy, University of Groningen).

Wśród najczęściej występujących autorów znaleźli się Mikko Tolonen (12) oraz Leo Lahti (12), czyli twórcy nazwy *bibliographic data science*. M. Tolonen jest historykiem, kierownikiem grupy badawczej Helsinki Computational History Group (COMHIS), zaś L. Lahti naukowcem związanym z *data science*. Prawie połowa publikacji w ramach BDS powstała z udziałem m.in. tych dwóch badaczy. To również autorzy afiliowani przy tej grupie badawczej stanowią największy odsetek (25%) w społeczności związanej z BDS. Można to interpretować z jednej strony, jako wyraz propagowania BDS przez członków tej grupy badawczej, z drugiej – jako dalsze rozwijanie tego podejścia badawczego.

Badania prowadzone pod szyldem BDS to w przeważającej większości wynik współpracy wielu badaczy. Ponad 90% tekstów to publikacje wieloautorskie. Co interesujące, model współpracy realizowany w ramach BDS w wielu przypadkach opiera się na połączeniu wiedzy dziedzinowej właściwej dla nauk humanistycznych z kompetencjami specjalistów od *data science* afiliowanych przy ośrodkach związanych z naukami ścisłymi lub technicznymi. Tabela 1 zawiera przykładowe publikacje wpisujące się w ten model współpracy.

Tabela 1

Model współpracy w ramach BDS

Publikacja	Autor	Afiliacja
Bryer, E., Rhujittawiwat, T., Comandur, S., Madrid, V., Riley, S., Rose, J., & Wilder, C. (2021, January). Analysis of Clustering Algorithms to Clean and Normalize Early Modern European Book Titles. In 2021 The 4th International Conference on Software Engineering and Information Management (pp. 106-112).	Evan Bryer; Theppatorn Rhujittawiwat; John R. Rose; Samyu Comandur; Vasco Madrid;	College of Engineering and Computing, University of South Carolina, United States
	Colin F. Wilder	Center for Digital Humanities, University of South Carolina, United States
	Stephanie M. Riley	College of Arts and Sciences, University of South Carolina, United States
Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., & Lahti, L. (2019). Interdisciplinary collaboration in studying newspaper materiality. In Twin Talks Workshop at DHN 2019 Proceedings of the Twin Talks Workshop at DHN 2019, co-located with Digital Humanities in the Nordic Countries (DHN 2019). CEUR-WS.org.	Eetu Mäkelä, Mikko Tolonen, Jani Marjanen, Annti Kanner, Ville Vaara	Helsinki Centre for Digital Humanities, University of Helsinki, Finland
	Leo Lahti	Department of Mathematics and Statistics, University of Turku, Finland
Péter, R., Szántó, Z., Seres, J., Bilicki, V., & Berend, G. (2020). AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts. In: Magyar Számítógépes Nyelvészeti Konferencia, (16). pp. 43-55. (2020)	Róbert Péter	University of Szeged, Institute of English and American Studies
	Zsolt Szántó, József Seres, Vilmos Bilicki, Gábor Berend	University of Szeged, Institute of Informatics
Pawłowski, A., Topolski, K., & Herden, E. (2021). Quantitative Analysis of Bibliographic Corpora : Statistical Features , Semantic Profiles , Word Spectra. In A. Pawłowski, J. Mačutek, S. Embleton, & G. Mikros (Eds.), Language and Text: Data, Models, Information, Applications (pp. 240-256). John Benjamins Publishing Company.	Adam Pawłowski, Elżbieta Herden	Institute of Information and Library Science, University of Wrocław, Poland
	Krzysztof Topolski	Institute of Mathematics, University of Wrocław, Poland

Model współpracy realizowany w ramach BDS opiera się więc na transferze metod i technik właściwych dla *data science* do subdyscyplin nauk humanistycznych, sprawiając, że ma charakter interdyscyplinarny, co jest typowe dla cyfrowej humanistyki.

Na obecnym etapie rozwoju BDS trudno jest jeszcze mówić o ośrodkach badawczych, ale należy odnotować aktywność badaczy afiliowanych przy wspomnianej wcześniej grupie badawczej COMHIS funkcjonującej w ramach Uniwersytetu w Helsinkach, Uniwersytecie Karoliny Południowej i Uniwersytecie Kalifornijskim (USA), Uniwersytecie w Segedynie (Węgry) czy Uniwersytecie Wrocławskim. To właśnie przy tych uczelniach afiliowani są autorzy prowadzący badania w ramach BDS.

WNIOSKI

Bibliographic data science interpretowane jako zastosowanie metod i technik *data science* w badaniach nad zasobami bibliograficznymi nie jest nowym obszarem badawczym, jednak osadzone w ramach cyfrowej humanistyki czyni je interesującym sposobem patrzenia na rzeczywistość, szczególnie historyczną, przez pryzmat metadanych dokumentów. To właśnie aspekt epistemiczny ma tutaj kluczowe znaczenie. BDS oferuje bowiem technologicznie zdeterminowane metodologie analizy i optymalizacji jakości metadanych jako narzędzia poznania rzeczywistości i tworzenia nowej wiedzy. Przy czym nie zrywa się tutaj z podejściem dedukcyjnym, tak jak w przypadku *big data* (Osika, 2020, s. 77), lecz konstruuje problemy badawcze i hipotezy osadzone w teoriach subdyscyplin cyfrowej humanistyki. Jest to szczególnie widoczne w warstwie ontologicznej. Problemy badawcze o charakterze metodologicznym, związane z oceną jakości metadanych w katalogach bibliotecznych i bibliografiach, stanowią tutaj odrębny nurt badań. Widać więc, że specyfika BDS jest nieco inna niż w przypadku „dziedzinowych" *data science* (np. *medical data science*, *biological data science*, *social data science*). Silny związek BDS z cyfrową humanistyką ujawnia się również w warstwie socjologicznej, o czym świadczą wyniki analizy społeczności skupionej wokół tego obszaru badawczego.

Na podstawie dotychczas przedstawionych rozważań można również poszukiwać związków BDS z informacją naukową i bibliotekoznawstwem. Biorąc pod uwagę, kto i jak uprawia BDS, wydaje się, że katalogi biblioteczne i bibliografie są tutaj traktowane przede wszystkim jako źródła danych badawczych bez większego angażowania przedstawicieli środowiska bibliotecznego. Skoro więc BDS wyciąga metadane poza ich docelowy kontekst społeczny, to *de facto* mamy do czynienia ze zjawiskiem zmiany przeznaczenia metadanych. I to właśnie na ten aspekt środowisko bibliotekarskie powinno zwrócić szczególną uwagę w projektowaniu oraz sposobach świadczenia usług informacyjnych. Chodzi tu-

taj nie tylko o możliwości, jakie dostarcza nam nowoczesna technologia w zakresie analizy i przetwarzania zasobów bibliograficznych, ale przede wszystkim o nowych interesariuszy zasobów bibliograficznych. Zasygnalizowane wcześniej związki między bibliometrią a BDS polegają na traktowaniu metadanych jako obiektów badawczych. W obydwu obszarach mamy do czynienia z zastosowaniem zaawansowanych technologii informacyjnych. BDS jest ewidentne osadzone w metodach i technologiach data science, ale w przypadku bibliometrii również można wskazać przykłady zastosowania metod uczenia maszynowego czy przetwarzania języka naturalnego (Abu-Jbara et al., 2013; Klein et al., 2021) w badaniach nad zasobami bibliograficznymi. To, co z pewnością różni te dwa obszary to cele i problemy badawcze oraz relacje z informacją naukową i bibliotekoznawstwem. Podstawowym zadaniem bibliometrii jest ocena stanu i rozwoju komunikacji piśmienniczej (głównie naukowej) z naciskiem na jej produktywność i efektywność, co sytuuję ją w polu badawczym informacji naukowej (Skalska-Zlat, 2017, s. 260). W przypadku BDS również mamy do czynienia z badaniem produkcji wiedzy (Lahti, Marjanen, et al., 2019, p. 6), jednak główna uwaga jest tutaj skupiona na determinantach społecznych i kulturowych oraz kontekście historycznym, co z kolei sytuuje BDS w polach badawczych subdyscyplin cyfrowej humanistyki. BDS dodatkowo eksponuje utylitarny i metodyczny aspekt w badaniach nad zasobami bibliograficznymi w postaci metod i technik analizy jakości metadanych realizowane z wykorzystaniem nowoczesnych technologii informacyjno-komunikacyjnych, co może być interesujące dla dostawców bibliograficznych usług informacyjnych w kontekście optymalizacji tworzonych i rozwijanych przez nich kolekcji.

BIBLIOGRAFIA

Abu-Jbara, Amjad; Ezra, Jefferson; Radev, Dragomir. (2013). Purpose and Polarity of Citation: Towards NLP-based Bibliometrics [online]. In: *Proceedings of NAACL-HLT*, pp. 596-606. [dostęp: 17.01.2022]. Dostępny w WWW: https://aclanthology.org/N13-1067.pdf

Bawden, David; Robinson, Lyn. (2015). *Introduction to Information Science*. Facet Publishing. https://doi.org/10.29085/9781783300761.

Bojar, Bożenna (Ed.). (2002). *Słownik encyklopedyczny informacji, języków i systemów informacyjno-wyszukiwawczych*. Warszawa: Wydaw. SBP.

Bryer, Evan; Rhujittawiwat, Theppatorn; Comandur, Samyu; Madrid, Vasco; Riley, Stephanie; Rose, John; Wilder, Colin. (2021). Analysis of Clustering Algorithms to Clean and Normalize Early Modern European Book Titles. *ACM International Conference Proceeding Series*, pp. 106-112. https://doi.org/10.1145/3451471.3451489.

Capurro, Rafael. (2010). Digital hermeneutics: An outline. *AI and Society*, 25(1), pp. 35-42. https://doi.org/10.1007/s00146-009-0255-9.

Ceusters, Werner. (2012). An Information Artifact Ontology Perspective on Data

Collections and Associated Representational Artifacts. *Studies in Health Technology and Informatics*, 180, pp. 68-72. https://doi.org/10.3233/978-1-61499-101-4-68.

Chakravartty, Anjan. (2004). Stance relativism: empiricism versus metaphysics. *Studies in History and Philosophy of Science Part A*, 35(1), pp. 173-184. https://doi.org/10.1016/j.shpsa.2003.12.002.

Czapnik, Grzegorz. (2016). Bibliomining w badaniach bibliotek cyfrowych. W: *Metody i narzędzia badań piśmiennictwa cyfrowego i jego użytkowników* pod red. Małgorzaty Góralskiej, Agnieszki Wandel. Wrocław: Wydaw. Uniwersytetu Wrocławskiego, s. 77-94.

Dalbello, Marija. (2011). A genealogy of digital humanities. *Journal of Documentation*, *67*(3), pp. 480-506. https://doi.org/10.1108/00220411111124550.

Dempsey, Lorcan. (2012). *Pretty interesting bibliographic data science role at Mendeley* [online]. Twitter; [dostęp: 17.01.2022]. Dostępny w WWW: https://twitter.com/lorcanD/status/190112947706662914.

Deng, Sai. (2010). Optimizing Workflow through Metadata Repurposing and Batch Processing. *Journal of Library Metadata*, *10*(4), pp. 219-237. https://doi.org/10.1080/19386389.2010.524862.

Eder, Maciej. (2014). Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii. *Teksty Drugie*, 2, s. 90-105.

Foulonneau, Muriel; Cole, Timothy. (2005). Strategies for Reprocessing Aggregated Metadata. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Ed. by Andreas Rauber, Stavros Christodoulakis, A Min Tjoa. Amsterdam: Springer, pp. 290-301. https://doi.org/10.1007/11551362_26.

Foulonneau, Muriel; Riley, Jenn. (2008). *Metadata for Digital Resources: Implementation, Systems Design and Interoperability*. Oxford: Chandos Publishing.

Gibert, Karina; Horsburgh, Jeffery; Athanasiadis, Ioannis; Holmes, Geoff. (2018). Environmental Data Science. *Environmental Modelling & Software*, 106, pp. 4-12. https://doi.org/10.1016/j.envsoft.2018.04.005.

Giudici, Paolo. (2018). Financial data science. *Statistics & Probability Letters*, 136, pp. 160-164. https://doi.org/10.1016/j.spl.2018.02.024.

Gkeredakis, Emmanouil; Fayard, Anne-Laure Fayard; Levina, Natalia. (2016). Data science as epistemic stance: advantages, risks and opportunities for the pursuit of knowledge. *Academy of Management Proceedings*, *1*. https://doi.org/10.5465/ambpp.2016.16538abstract.

Hjørland, Birger. (2002). Epistemology and the socio-cognitive perspective in information science. *Journal of the American Society for Information Science and Technology*, 53(4), pp. 257-270. https://doi.org/10.1002/asi.10042.

Hjørland, Birger. (2017). Domain Analysis. *Knowledge Organization*, 44(6), pp. 436-464. https://doi.org/10.5771/0943-7444-2017-6-436.

Hjørland, Birger; Albrechtsen, Hanne. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, *46*(6), pp. 400-425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hjørland, Birger; Hartel, Jenna. (2003). Afterword: Ontological, epistemological and sociological dimensions of domains. *Knowledge Organization*, 30(3-4), pp. 239-245.

Hripcsak, George; Duke, Jon; Shah, Nigam; Reich, Christian; Huser, Vojtech; Schuemie, Martijn; Suchard, Marc; Park, Rae Woong; Wong, Ian Chi Kei; Rijnbeek, Peter; van der Lei, Johan; Pratt, Nicole; Norén, Niklas; Li, Yu-Chuan; Stang, Paul; Madigan, David; Ryan, Patrick. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, 216, pp. 574-578.

Klein, Jennifer; Baker, Nancy; Foil, Nancy; Zorn, Kimberley; Urbina, Fabio; Puhl, Ana; Ekins, Sean. (2021). *Using Bibliometric Analysis and Machine Learning to Identify Compounds Binding to Sialidase-1*. ACS Omega, 6(4), 3186-3193. https://doi.org/10.1021/acsomega.0c05591

Lahti, Leo; Ilomäki, Niko; Tolonen, Mikko. (2015). A quantitative study of history in the english short-title catalogue (ESTC), 1470-1800. *LIBER Quarterly*, 25(2), pp. 87-116. https://doi.org/10.18352/lq.10112.

Lahti, L., Mäkelä, Eetu; Tolonen, Mikko. (2020). Quantifying bias and uncertainty in historical data collections with probabilistic programming [online]. *CEUR Workshop Proceedings*, *2723*, pp. 280-289. [dostęp: 17.01.2022]. Dostępny w WWW: http://ceur-ws.org/Vol-2723/short46.pdf.

Lahti, Leo; Marjanen, Jani; Roivainen, Hege; Tolonen, Mikko. (2019). Bibliographic data science and the history of the book (C. 1500-1800). *Cataloging and Classification Quarterly*, 57(1), pp. 5-23. https://doi.org/10.1080/01639374.2018.1543747.

Lahti, Leo; Vaara, Ville; Marjanen, Jani; Tolonen, Mikko. (2019). Best Practices in Bibliographic Data Science [online]. In: *Proceedings of the Research Data And Humanities (RDHUM) 2019 Conference: Data, Methods And Tools. Studia humaniora Ouluensia,* vol. 17. Ed. by Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen, Katja Västi. University of Oulu, pp. 57-65. [dostęp: 17.06.2022]. Dostępny w WWW: https://helda.helsinki.fi/handle/10138/310192.

Lüschow, Andreas; Tello, Jose Calvo. (2021). Towards genre classification in the library catalog [online]. *CEUR Workshop Proceedings*, *2836*. [dostęp: 17.01.2022]. Dostępny w WWW: http://ceur-ws.org/Vol-2836/qurator2021_paper_9.pdf.

Marjanen, Jani; Vaara, Ville; Kanner, Antti; Roivainen, Hege; Mäkelä, Eetu; Lahti, Leo; Tolonen, Mikko. (2019). A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771-1917. *Journal of European Periodical Studies*, 4(1), pp. 54-77. https://doi.org/10.21825/jeps.v4i1.10483.

Mayernik, Matthew. (2019). Metadata accounts: Achieving data and evidence in scientific research. *Social Studies of Science*, 49(5), pp. 732-757. https://doi.org/10.1177/0306312719863494.

Moretti, Franco. (2013). *Distant Reading*. London: Verso.

Nicholson, Scott. (2011). Bibliomining for Library Decision-Making. *Encyclopedia of Data Warehousing and Mining, Second Edition*. https://doi.org/10.4018/9781605660103.ch025.

Nicholson, Scott; Hwang, San-Yih; Keezer, Paula; O'Neill, Edward. (2003). The bibliomining process: Data warehousing and data mining for libraries. *Proceedings of the ASIST Annual Meeting*, 40, pp. 478-479. https://doi.org/10.1002/meet.1450400184.

Nowak, Adam. (2016). Bibliografia a katalog – dyskusja o pojęciach i terminach. Historyczny zarys problematyki. *Przegląd Biblioteczny*, 84(1), pp. 5-26. https://doi.org/https://doi.org/10.36702/pb.472.

Olson, Hope. (1996). Dewey Thinks Therefore He Is: The Epistemic Stance of Dewey and DDC. *Knowledge Organization and Change. Proceedings of the Fourth International ISKO Conference 15-18 July 1996, Washington, D.C*, 5(1995), pp. 302-312.

Osika, Grażyna. (2020). Datafikacja – implikacje epistemologiczne. *Przegląd Filozoficzny*, 3(115), s. 71-85. https://doi.org/10.24425/pfns.2020.133975.

Pawłowski, Adam; Herden, Elżbieta; Walkowiak, Tomasz. (2021). Book Genre and Author s Gender Recognition Based on Titles : the Example of the Bibliographic Corpus of Microtexts. In: *Language and Text: Data, Models, Information, Applications.* Ed. by Adam Pawłowski, Jan Mačutek, Shella Embleton, George Mikros. Amsterdam: John Benjamins Publishing Company, pp. 226-237.

Pawłowski, Adam; Topolski, Krzysztof; Herden, Elżbieta. (2021). Quantitative Analysis of Bibliographic Corpora : Statistical Features , Semantic Profiles, Word Spectra. In: Language and Text: Data, Models, Information, Applications. Ed. by Adam Pawłowski, Jan Mačutek, Shella Embleton, George Mikros. Amsterdam: John Benjamins Publishing Company, pp. 240-256.

Péter, Róbert; Szántó, Zsolt; Seres, József; Bilicki, Vilmos; Berend, Gábor. (2020). AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts [online]. SZTE Repository of Papers and Books [dostęp: 17.01. 2022]. Dostępny w WWW: http://acta.bibl.u-szeged.hu/67682/

Sadowska, Jadwiga. (2018). Z problemów urzędowej rejestracji druków w II Rzeczypospolitej: egzemplarz obowiązkowy, bibliografia narodowa, statystyka wydawnicza. *Roczniki Biblioteczne*, 61, s. 191-206. https://doi.org/10.19195/0080-3626.61.9.

Skalska-Zlat, Marta. (2017). Bibliometria. W: Marta Skalska-Zlat & Anna Żbikowska-Migoń (Eds.), *Encyklopedia Książki* (Vol. 1, s. 258-260). Wrocław: Wydaw. Uniwersytetu Wrocławskiego.

Semeler, Alexandre Ribas; Pinto, Adilson Luiz; Rozados, Helen Beatriz Frota. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science*, 51(3), pp. 771-780. https://doi.org/10.1177/0961000617742465.

Sosińska-Kalata, Barbara. (2018). Big data (dane masowe) w nauce o informacji. *Zagadnienia Informacji Naukowej – Studia Informacyjne*, 112(2), s. 7-35. https://doi.org/10.36702/zin.390

Sosińska-Kalata, Barbara; Roszkowski, Marcin. (2016). Organizacja informacji i wiedzy. W: *Nauka o informacji* pod red. Wiesława Babika. Warszawa: Wydaw. SBP., s. 305-358.

Tolonen, Mikko; Marjanen, Jani; Roivainen, Hege; Lahti, Leo. (2019). Scaling up bibliographic data science [online]. *CEUR Workshop Proceedings*, *2364*, pp. 450-456. [dostęp: 17.01.2022]. Dostępny w WWW: http://ceur-ws.org/Vol-2364/41_paper.pdf.

Underwood, Ted. (2020). Machine Learning and Human Perspective. *PMLA/Publications of the Modern Language Association of America*, 135(1), pp. 92-109. https://doi.org/10.1632/pmla.2020.135.1.92

Woźniak-Kasperek, Jadwiga. (2015). Bibliografia a katalog biblioteczny – dyskusja o pojęciach i terminach. *Przegląd Biblioteczny*, 83(4), s. 517-532. https://doi.org/10.36702/pb.513.

Artykuł w wersji poprawionej wpłynął do Redakcji 17 lutego 2022 r.

MARCIN ROSZKOWSKI
Faculty of Journalism, Information and Book Studies
University of Warsaw
e-mail: m.roszkowski@uw.edu.pl
ORCID 0000-0001-7396-4685

BIBLIOGRAPHIC DATA SCIENCE – CONCEPTUALIZATION OF THE RESEARCH AREA

KEYWORDS: Bibliographic data science. Digital humanities. Library catalogs. Bibliographic databases. Data harmonization.

ABSTRACT: **Thesis/Objective** – The subject of the article is a new research area called bibliographic data science, which is understood as the use of data science methods and technologies in the research on the content of library catalogs and bibliographies. **Research methods** – The methodological framework for this study is based on qualitative analysis employing critical literature review and domain analysis. **Results** – Bibliographic data science is focused on a pragmatic approach to data-supported research on bibliographic resources within the field of digital humanities. Issues researched in this area are related to respective disciplines and subfields that constitute digital humanities as well as methods of metadata quality optimization and harmonization. Strong relationships between bibliographic data science and digital humanities are prominent within the scholarly community growing around this research area.