Big data (dane masowe) w nauce o informacji

Barbara Sosińska-Kalata
Katedra Informatologii, Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski

Abstrakt

Cel/Teza: Celem artykułu jest omówienie głównych cech zjawiska określanego mianem big data, jego znaczenia dla problematyki badawczej nauki o informacji oraz próba wstępnej oceny stopnia zainteresowania nim badaczy tej dyscypliny.
Koncepcja/Metody badań: Krytyczna analiza piśmiennictwa przedmiotu wykorzystana została do omówienia istoty zjawiska big data oraz związanych z nim zmian w modelu badań naukowych, który w coraz większym zakresie znajduje zastosowanie w różnych dziedzinach współczesnej nauki. Rosnącą popularność badań big data w nauce zilustrowano wynikami analizy bibliometrycznej piśmiennictwa zarejestrowanego w interdyscyplinarnej bazie Scopus. Ocenę stopnia zainteresowania problematyką big data w nauce o informacji oparto na bibliometrycznej analizie piśmiennictwa indeksowanego w dziedzinowej bazie EBSCO – Library and Information Science and Technology Abstracts (LISTA).
Wyniki i wnioski: Zagadnienie big data można traktować jako kolejną fazę rozwoju technologii komputerowej i jej zastosowań w różnych dziedzinach nauki i praktyki. W środowisku wielkich zasobów danych zapisanych w cyfrowym formacie, technologie big data zapewniają wgląd w wiedzę, której nie można byłoby wydobyć tradycyjnymi metodami wyszukiwania informacji. W tym sensie technologie te wspierają procesy transferu wiedzy między ludźmi, które stanowią główny przedmiot zainteresowań nauki o informacji. Analiza piśmiennictwa indeksowanego w bazie LISTA potwierdziła, że rozwój technologii big data i jej zastosowań stanowi istotne wyzwanie dla nauki o informacji, którym zainteresowanie badaczy systematycznie rośnie, jakkolwiek nie jest ono jeszcze w tej dyscyplinie bardzo duże. Analiza tematyki tego piśmiennictwa potwierdziła też, że problematyka big data łączy się z kluczowymi obszarami badań nauki o informacji. Badania dotyczące big data najczęściej prezentowane są na łamach czasopism specjalizujących się w ilościowych badaniach informacji (bibliometrii, naukometrii, altmetrii), informatyce medycznej, problematyce systemów informacyjnych i wyszukiwania informacji oraz w zarządzaniu informacją. W czasopismach o szerokim profilu tematycznym obejmującym całe pole badawcze nauki o informacji publikacje na temat big data dotychczas ukazywały się rzadko. Autorami największej liczby artykułów dotyczących tej problematyki są badacze związani z ośrodkami naukowymi w Stanach Zjednoczonych, w Wielkiej Brytanii i w Chinach. Piśmiennictwo dotyczące badań big data w nauce o informacji charakteryzuje duża różnorodność podejmowanej tematyki szczegółowej. Dominuje tematyka należąca do obszaru nauk komputerowych oraz mediów społecznych, ale do zagadnień często omawianych należą też metadane, zarządzanie i dzielenie się wiedzą, biblioteki cyfrowe, bibliometria oraz kwestie związane z informatyką medyczną i ochroną zdrowia.
Ograniczenia badań: Omówione badanie ma charakter sondażowy i przeprowadzone zostało na indeksowanym w bazie LISTA piśmiennictwie, w którego opisie tematycznym użyty został termin „big data". Piśmiennictwo prezentujące problematykę związaną z badaniem wielkich zbiorów danych, w którego indeksowaniu nie użyto tego terminu, nie zostało zatem uwzględnione w badaniu. Ponadto polityka indeksowania bazy LISTA, w szczególności względnie mała reprezentacja czasopism wydawanych w innych językach niż angielski wśród indeksowanych w niej źródeł, może ograniczać reprezentatywność uzyskanych wyników dla badań dotyczących big data, związanych z problematyką nauki o informacji, w skali globalnej.

Oryginalność/Wartość poznawcza: Zgodnie z wiedzą autorki, artykuł jest pierwszą próbą oceny stopnia zainteresowania problematyką big data w nauce o informacji.

Słowa kluczowe

Badanie bibliometryczne. Big data. Dane masowe. Nauka o informacji. Problematyka badawcza.

Otrzymany: 11 lutego 2018. Zrecenzowany: 23 lutego 2018. Zaakceptowany: 5 marca 2018.

1. Wprowadzenie

Dane masowe, obecnie najczęściej określane angielskim terminem „big data", oznaczają zbiory danych, które przyrastają w sposób nieograniczony, dla których pamięć musi być rezerwowana dynamicznie i których nie da się przetwarzać metodami tradycyjnymi. Dane masowe występują zatem w nadzwyczaj dużych ilościach, liczonych w petabajtach (PB), zettabajtach (ZB), czy nawet jottabajtach[1], a to implikuje konieczność nowego podejścia do ich gromadzenia, magazynowania, przetwarzania i transmisji. Według raportu firmy Intel, o produkcji big data można mówić wtedy, gdy organizacja generuje medianę 300 terabajtów danych tygodniowo (Intel, 2012). Należy dodać, że szacunki te odnoszą się przede wszystkim do wielkich zasobów danych wykorzystywanych w analizach wspierających procesy decyzyjne we współczesnym marketingu i zarządzaniu biznesem. Wielkie ilości danych generuje też nowoczesna aparatura naukowa i technologicznie zaawansowane narzędzia, np. jedno doświadczenie przeprowadzone w CERN-ie przy użyciu Wielkiego Zderzacza Hadronów generuje około 40 terabajtów (TB) danych w ciągu 30 minut, a jeden przelot odrzutowca dostarcza około 10 TB danych (Jacobfeuerborn, 2013). Wielkie zasoby danych tworzą ludzie, publikując w przestrzeni cyfrowej różnego typu komunikaty i pozostawiając ślady swojej aktywności w sieciach komputerowych, np. w postaci logów do różnych stron internetowych czy też kwerend kierowanych do wyszukiwarek. W 2012 r. na świecie opublikowanych zostało ponad 1,57 mln artykułów naukowych, co oznacza, że na każdą minutę tego roku przypadały 3 nowe artykuły (Ferstein, 2012). Już w 2009 r. Facebook informował, że spółka dysponuje 1 PB danych, natomiast w 2016 r. w posiadaniu Google było 15 eksabajtów[2] danych (Patgiri & Ahmed, 2016). Obecnie Google przetwarza 40 tys. pytań w każdej sekundzie, tj. 3,5 miliarda pytań dziennie[3]. Codziennie ogromne zasoby danych generuje sieć powiązanych z sobą milionów inteligentnych urządzeń współtworzących dynamicznie rozwijający się Internet Rzeczy (ang. *Internet of Things*, IoT).

Przytoczone przykłady dają wyobrażenie o wielkości zasobów określanych mianem „big data". W różnych dziedzinach skala wielkości zasobów przetwarzanych metodami big data może być inna, zawsze jednak mówić będziemy o wielkościach znacząco większych niż te, które stanowiły podstawę analiz prowadzonych metodami tradycyjnymi. Zasoby big data mogą być zarówno ustrukturyzowane (jak np. dane o transakcjach biznesowych, przechowywane w relacyjnych bazach danych, czy też dane gromadzone w wielkich bazach bibliograficznych, patentowych, rejestrów medycznych itp.), częściowo ustrukturyzowane

[1] Petabajt (PB) jest jednostką używaną do oznaczania biliarda bajtów, w których mierzona jest pojemność największych pamięci masowych; 1 PB jest równy 1015 bajtów, tj. ok. 1020 TB, zettabajt (ZB) oznacza tryliard bajtów, tj. 1021 bajtów, a jottabajt (YT) to kwadrylion bajtów, czyli 1024 bajtów.

[2] Eksabajt (EB) jest równy 1018 bajtów, tj. ok. 1 040 816 TB.

[3] Dane za Google Search Statistics (http://www.internetlivestats.com/google-search-statistics/).

(np. pełne teksty opatrzone tagami), jak i nieustrukturyzowane (np. wiadomości e-mail czy komentarze generowane w mediach społecznościowych).

Dla big data wielkość zasobów jest kluczowa, jednak wielu badaczy przekonuje, że nie tylko o ich wielkość chodzi. Na przykład, za cechę równie istotną dla zasobów big data uznaje się ich heterogeniczność, wielką różnorodność formatów i sposobów reprezentacji danych. Pogląd ten znalazł odzwierciedlenie w jednej z najpopularniejszych do dziś definicji „big data", którą w 2001 r. sformułował Doug Laney, odwołując się do trzech atrybutów, które uznał za konstytutywne dla tego typu zasobów i których nazwy zaczynają się na literę „v", tj. do tzw. formuły „3Vs": *volume* (wielkość), *velocity* (szybkość), *variety* (różnorodność); (Laney, 2001). Jednak dwanaście lat później, jak twierdzi Alon Friedman, w 2013 r. Laney zrewidował swoją głośną definicję, stwierdzając, iż:

> Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization (Friedman, 2017, 135).

Zatem big data charakteryzuje przede wszystkim wielkość zasobów, której towarzyszy albo szybki przyrost, albo duża różnorodność, albo obie te własności równocześnie. Nadzwyczajna wielkość zasobów big data w powiązaniu z szybkością ich narastania i różnorodnością formatów i reprezentacji danych czynią te zasoby zbyt złożonymi, aby można było je magazynować i przetwarzać tradycyjnymi metodami.

W miarę wzrostu zainteresowania problematyką big data kolejni badacze proponowali doprecyzowanie definicji „3Vs", wskazując inne atrybuty szybko rosnących zasobów danych, które uznawano za specyficzne dla nich i których nazwy, wzorem definicji Laneya, zaczynają się od litery „v": *variability* (zmienność), *veracity* (wiarygodność), *value* (wartość), *validity* (ważność), *volatility* (ulotność), *virtual* (faktyczność), *visualization/visibility* (wizualizacja/widoczność), *vitality* (witalność), *vaccum* (próżnia) (por. Patgiri & Ahmed, 2016). Dobieranie kolejnych v-atrybutów charakteryzujących big data stało się popularne w piśmiennictwie poświęconym temu zjawisku, jednak za faktycznie najistotniejsze jego cechy uznać należy przede wszystkim *volume* (wielkość) i *complexity* (złożoność).

Przetwarzanie zasobów big data wymaga stosowania nowej technologii i nowych metod analitycznych. Podstawą nowego podejścia do przetwarzania wielkich zasobów danych jest założenie, że jego głównym celem jest wydobywanie z tych zasobów ukrytej w nich nowej wiedzy przez stosowanie metod i technik określanych ogólnie analityką danych (ang. *data analytics*). Według Google Trends, *data analytics* jest terminem wyszukiwawczym najsilniej powiązanym z tematem big data. Przeprowadzona przez Jonathana Stuarta Warda i Adama Barkera analiza różnych definicji big data wykazała, że intensywnie rozwijane narzędzia i metody analizy danych masowych stanowią trzeci, konstytutywny składnik zjawiska big data (Ward & Barker, 2013). Do podobnych wniosków doszli też Andrea De Mauro, Marco Greco i Michele Grimaldi, analizując tematykę piśmiennictwa dotyczącego big data, które zarejestrowano w interdyscyplinarnej bazie Scopus (De Mauro et al., 2016). Metody eksploracji i analizy danych masowych ukierunkowane są na szukanie relacji i wzorów powiązań między danymi oraz szacowanie prawdopodobieństwa ich występowania, co pozwala następnie przewidywać trendy oraz rekomendować działania, decyzje i innowacje optymalne w określonych sytuacjach, w odniesieniu do potrzeb określonej grupy klientów i określonych celów. Technologie big data otworzyły nowe możliwości

zdobywania wiedzy niezwykle użytecznej dla nowoczesnego biznesu, gdzie od kilku lat są intensywnie wykorzystywane. Technologie te dostarczyły także nauce nowych, potężnych narzędzi badawczych, zapewniających znacznie bardziej niż dotąd szczegółowy wgląd w rozmaite zjawiska i procesy naturalne, techniczne i społeczne, a także w słabo dotąd poznane własności ludzkiej twórczości.

2. Big data i analityka danych masowych w nauce, w naukach społecznych i humanistyce

Informatyzacja aparatury badawczej, cyfryzacja informacji i wiedzy o człowieku i otaczającym go świecie, która obejmuje coraz większe obszary naszego życia, oraz wielki wzrost mocy obliczeniowej komputerów zmieniają sposób uprawiania nauki w niemal wszystkich już dziedzinach, oferując nowe podejście, nowe narzędzia i nowe metody poznawania świata i rozwiązywania problemów. Zjawisko to, zapoczątkowane w latach 90. XX w. intensywnym rozwojem technologii sieciowych, technik *data-mining* i *cloud computing* oraz ich wykorzystaniem w genetyce i astronomii, zostało określone mianem Czwartego Paradygmatu w ewolucji nauki (Hey et al., 2009; Jacobfeuerborn, 2013). Według Jima Graya, badacza z laboratoriów Microsoftu, który dziesięć lat temu określenie to zaproponował, trzy pierwsze podstawowe paradygmaty w rozwoju nauki stanowiły najdawniejszy paradygmat empiryczny, oparty na opisie zjawisk naturalnych, zapoczątkowany przez prace Galileusza, Johanna Keplera czy Tycho de Brahe'a paradygmat teoretyczny, oparty na modelowaniu zjawisk teoretycznej generalizacji oraz rozwijany w ostatnich kilkudziesięciu latach paradygmat komputacyjny, oparty na komputerowej symulacji złożonych zjawisk. Czwarty paradygmat, charakteryzujący e-naukę, oparty jest na intensywnym wykorzystywaniu danych cyfrowych w badaniach naukowych. Dane pozyskiwane są przez aparaturę badawczą lub generowane przez symulatory, a następnie przetwarzane przez oprogramowanie komputerowe; informacja i wiedza przechowywane są w pamięciach komputerowych; badacz analizuje zawartość baz danych czy plików komputerowych, korzystając z metod statystycznych i narzędzi zarządzania danymi. W ten sposób cały cykl badawczy oparty zostaje na cyfrowych danych reprezentujących badany świat oraz procesach ich komputerowego przetwarzania.

Viktor Mayer-Schönberger i Kenneth Cukier piszą, że technologie big data zmieniają nasze myślenie, pracę i życie (Mayer-Schönberger & Cukier, 2014). Wykorzystując metody eksploracji danych i wyodrębniania wzorów powiązań w wielkiej skali, zapewnianej przez dane masowe, przed nauką otworzyły one nowy sposób poznawania świata, który opiera się na zastąpieniu modelu wyjaśniania badanych zjawisk i procesów przez ustalanie ich przyczyn, modelem ustalania korelacji między szczegółowymi danymi opisującymi (reprezentującymi) te zjawiska i procesy oraz ich kontekst (środowisko, sytuacje) w przestrzeni cyfrowej. Przyjmuje się założenie, że ustalenie korelacji między elementami wielkich zbiorów danych wystarcza do uzyskania nowej wiedzy, poznania nieznanych dotąd własności, procesów i aspektów naszej rzeczywistości. Korelacje te nie muszą wyjaśniać, dlaczego coś się dzieje, ale informują, że to się dzieje, pozwalając przewidywać kierunki i sposoby rozwoju badanych zjawisk. Jak piszą Mayer-Schönberger i Cukier:

> Nie zawsze musimy znać przyczyny jakiegoś zjawiska, możemy po prostu pozwolić danym przemawiać za siebie (Mayer-Schönberger & Cukier, 2014, 30).

Zwiększenie ilości danych, na których przeprowadzane są analizy, umożliwia odkrywanie ukrytych powiązań i modelowanie ich powtarzalnych schematów, których dostrzeżenie było niemożliwe przy mniejszej ilości informacji. Technologie big data w niektórych przypadkach umożliwiają poddanie analizie wszystkich danych, które dotyczą badanych zjawisk, a nie tylko np. ich próby losowej, którą operuje się w tradycyjnych badaniach ilościowych, zakładając (starając się zapewnić) jej reprezentatywność. Znacznie większa szczegółowość i kompletność danych stanowiących podstawę poznania naukowego równocześnie umożliwia jego wielką skalowalność – od poznawania ogólnego kierunku do poznawania najbardziej szczegółowych detali. Charakterystyczną cechą nowego podejścia do poznawania świata jest też rezygnacja z zachowania dużej dokładności dokonywanych pomiarów, którą w wielu przypadkach może zastąpić lepsze zrozumienie badanych zjawisk dzięki wykorzystaniu wielkiej liczby różnorodnych danych opisujących te zjawiska.

Niektórzy badacze uważają, że model badania oparty na wykorzystaniu big data zapewnia też większy obiektywizm poznania niż badania oparte na formułowaniu założeń i hipotez, budowaniu teorii oraz ich weryfikacji na podstawie relatywnie małych prób badawczych. Wielką dyskusję w nauce wywołał Chris Anderson, redaktor naczelny magazynu *Wired*, ogłaszając nawet „koniec teorii" i twierdząc, że w epoce big data formułowanie spekulatywnych teorii wyjaśniających, czym są analizowane dane, jest niepotrzebne, wystarczy bowiem poznanie korelacji, w które te dane wchodzą (Anderson, 2008). Nie ulega wątpliwości, że identyfikacja korelacji zachodzących w wielkich zbiorach danych pozwala na stawianie nowych pytań, otwierając nowe możliwości poznawcze. Niemniej jednak, zarówno teza o obiektywizmie badań opartych na analizie wielkich zasobów danych, jak i teza o końcu teorii nie są przekonujące. W szczególności trzeba podkreślić, że chociaż nadzwyczajna wielkość analizowanych zbiorów danych zmniejsza ryzyko błędu związanego z pominięciem danych istotnych, to jednak wiarygodność wyników analiz metodami big data zawsze zależy od jakości danych poddawanych analizie, a ta z kolei – od metod pozyskiwania danych, wykorzystanych źródeł, a także przygotowania danych do analizy. Poglądowi o obiektywizmie badań opartych na zasobach big data i ilościowych metodach ich analizy przeczy też fakt, iż zakres gromadzenia danych (np. przez określone urządzenia pomiarowe) i interpretacja wyników statystycznych analiz wielkich zbiorów danych zależne są od instrumentarium i celu prowadzonych badań.

Niezależnie od kontrowersji dotyczących oceny stopnia wpływu zjawiska big data na transformację badań naukowych nie podlega dyskusji to, że wpływ ten jest coraz silniejszy i dotyczy coraz większej liczby dyscyplin. Kluczową rolę odgrywa tu zmiana nastawienia do tego, jak dane mogą być wykorzystywane, która nastąpiła w ostatnich kilkunastu latach. Po pierwsze, wraz z przekonaniem, że przydatność danych nie kończy się z chwilą osiągnięcia celu, dla którego były one gromadzone (np. do wykonania pewnego działania), upowszechniła się archiwizacja danych; wtórne wykorzystanie danych staje się źródłem inspiracji i innowacji. Po drugie, dostępność technologii cyfrowych i coraz szerszy zakres ich zastosowania implikują tzw. danetyzację rzeczywistości, tj. zbieranie danych o wszystkim, w tym o kwestiach, o których dotąd nie myślano jako o źródłach danych (np. naprężenia w konstrukcji mostu, wibracje silnika, miejsce przebywania konkretnej osoby, logi do serwisów internetowych, słowa wpisywane przez użytkowników do wyszukiwarek internetowych) i przetwarzanie ich w kwantyfikowalny format.

Wielkie możliwości wykorzystania nowego modelu badań naukowych przed naukami społecznymi otworzyły w szczególności media społecznościowe oraz masowe generowanie,

gromadzenie i przetwarzanie danych o ludzkich zachowaniach zbiorowych i indywidualnych. Big data staje się coraz powszechniejszym mechanizmem, którym ludzie posługują się, nadając sens otaczającej ich rzeczywistości i równocześnie dostarczając niezmiernie bogaty materiał badawczy (Klous & Wielaard, 2016). Z kolei masowa digitalizacja dziedzictwa kulturowego i nowe instrumentarium badań ludzkiej twórczości dały podstawę do coraz szerszego zastosowania metod big data w humanistyce.

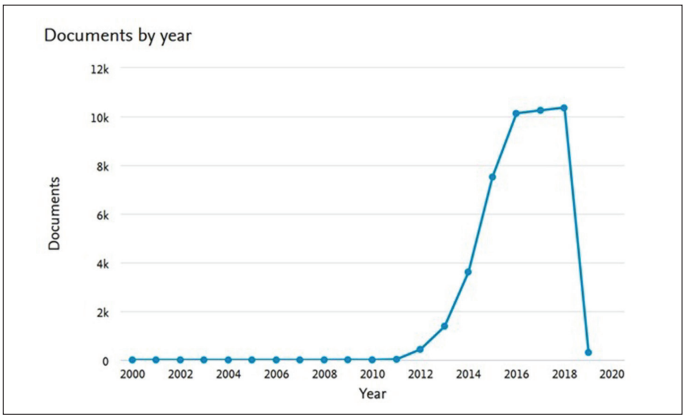
Szybko rosnącą popularność wykorzystywania danych masowych w badaniach naukowych oraz dystrybucję tego typu badań w różnych dziedzinach nauki ilustrują wyniki prostego badania bibliometrycznego, przeprowadzonego na interdyscyplinarnej bazie Scopus. Baza ta, utworzona przez wydawnictwo Elsevier w 2004 r., rejestruje piśmiennictwo ze wszystkich obszarów nauki od 1970 r., w tym artykuły opublikowane w ponad 22,8 tys. czasopismach naukowych i 8,3 mln artykułów opublikowanych w materiałach konferencyjnych. Obecnie baza Scopus zawiera ponad 71 mln rekordów (Scopus, 2018).

Na podstawie wyszukiwania za pomocą terminu „big data" przeprowadzonego 18 listopada 2018 r. w polu słów kluczowych, w którym umieszczane są terminy wyodrębnione jako charakterystyczne dla tematyki omawianej w indeksowanym dokumencie, z zasobów bazy Scopus wyodrębniono zbiór piśmiennictwa dotyczącego zjawiska big data. Ponieważ celem wyszukiwania było wyodrębnienie wszystkich publikacji na ten temat, niezależnie od ich formy i przynależności dziedzinowej, nie zastosowano ograniczeń formalnych ani dziedzinowych. W rezultacie otrzymano 44 052 rekordów publikacji wydanych w okresie od 2000 do 2019 r.[4]. Analiza rozkładu chronologicznego otrzymanego zbioru rekordów wykazała, że w Scopus zarejestrowano jedną publikację wydaną w 2000 r., a w latach 2001–2010 – od jednej do siedmiu publikacji. Wyraźny wzrost publikacji dotyczących big data następuje w 2011 r., w którym zarejestrowano już 25 dokumentów o tej tematyce, z czego 21 opublikowanych zostało w czasopismach i materiałach konferencyjnych z zakresu informatyki. W kolejnych latach wzrostowa tendencja utrzymuje się, a tempo wzrostu szybko rośnie: w 2012 r. zarejestrowano 435 publikacji o big data, w 2013 – 1374, w 2014 – 3618, w 2015 – 7536 publikacji. W 2016 r. liczba publikacji dotyczących big data przekroczyła 10 tys., a tempo jej wzrostu w kolejnych latach znacznie zmniejszyło się, co świadczyć może o pewnej stabilizacji aktywności środowiska badawczego zajmującego się tym zagadnieniem. W 2018 r. do połowy listopada zarejestrowano 10 369 publikacji oraz 306, które ukazały się już z datą wydania 2019 (Rys. 1).

Pierwszą publikacją dotyczącą problematyki big data, którą zarejestrowano w Scopus jest artykuł zakwalifikowany zarówno do nauk komputerowych, jak i nauk społecznych, dotyczący metod kompresji obrazów, wydany w 2000 r. w czasopiśmie *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS Archives* (Rys. 2).

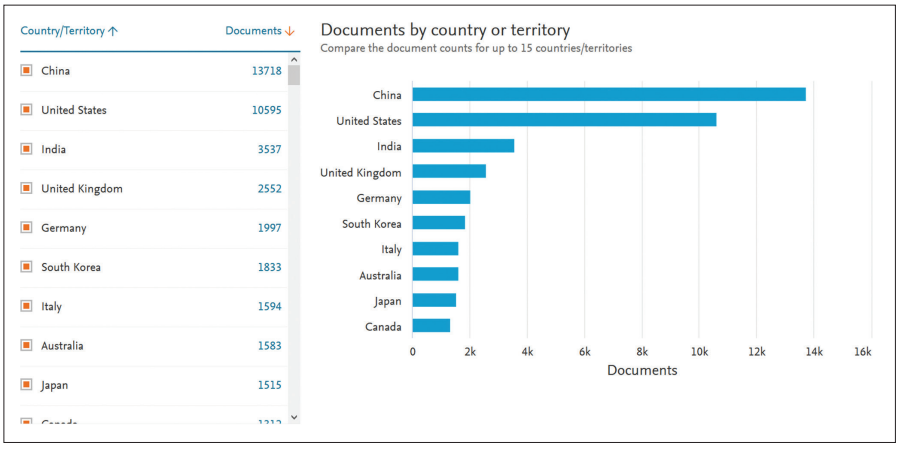
Rozkład geograficzny zbioru publikacji o big data wyodrębnionego z bazy Scopus prezentuje wyraźną dominację w badaniu tego zagadnienia dwóch krajów: Chin i Stanów Zjednoczonych (Rys. 3). Na chińskie ośrodki badawcze przypada 31% badań omówionych w tym piśmiennictwie, a na USA 24%. Dwoma kolejnymi państwami, w których prowadzona jest największa liczba badań o tej tematyce, są Indie (8%) i Wielka Brytania (ok. 6%).

[4] Rejestracja artykułów z rokiem wydania 2019 wynika z coraz powszechniejszej tendencji przyspieszania publikacji online tekstów zaakceptowanych i przygotowanych do druku w numerach, które formalnie często ukazać się mają dopiero za kilka miesięcy.



Rys. 1. Chronologiczny rozkład publikacji dotyczących big data, zarejestrowanych w bazie Scopus [wyszukiwanie: 18.11.2018]

Rys. 2. Rekord pierwszego artykułu dotyczącego problematyki big data, zarejestrowanego w Scopus [18.11.2018]

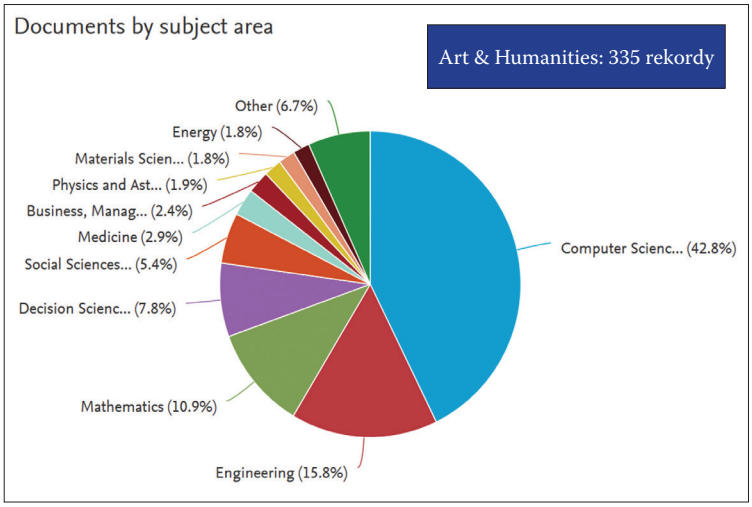


Rys. 3. Rozkład geograficzny publikacji dotyczących big data zarejestrowanych w bazie Scopus [18.11.2018]

Rozkład piśmiennictwa według dyscyplin naukowych uwidacznia jednoznaczną dominację informatyki w badaniach dotyczących big data: przypada na nią blisko 43% wszystkich publikacji (Rys. 4). Big data to przede wszystkim zagadnienie informatyczne, ale też związane z badaniami zarówno podstawowymi, jak i aplikacyjnymi w innych dziedzinach. W świetle danych otrzymanych w wyniku przeprowadzonego wyszukiwania dziedzinami wyodrębnionymi według kategoryzacji stosowanej w Scopus, w których zagadnienie big data zyskało znaczącą popularność, są: nauki techniczne (15,8%), matematyka (10,9%) i tzw. *decision sciences* – interdyscyplinarne pole badań zajmujące się wykorzystywaniem technik ilościowych w podejmowaniu decyzji w zarządzaniu i biznesie (7,8%). W dwóch pierwszych przypadkach mamy do czynienia z technicznymi aspektami przetwarzania danych masowych, organizacji i realizacji tego procesu, w drugim natomiast z matematycznymi aspektami tworzenia algorytmów wykorzystywanych w przetwarzaniu big data. *Decision sciences* zajmują się działalnością, w której technologie big data znajdują najszersze zastosowanie. Czwartą kategorię dziedzinową pod względem liczby publikacji o big data stanowią nauki społeczne, na które przypada 5,4% publikacji (4319 rekordów). Trzeba jednak zaznaczyć, że z kategorii tej w Scopus wyłączone zostały zarówno wspomniane *decision sciences*, jak i nauki o zarządzaniu i biznesie (1917 rekordów, 2,4%).

Analizując rozkład piśmiennictwa o big data według kryterium dziedzinowego, warto zwrócić uwagę na nieobecność humanistyki i nauk o sztuce wśród kategorii wyróżnionych przez narzędzia analityczne Scopus w uzyskanym wyniku wyszukiwania. Spośród publikacji z tego zakresu w Scopus zarejestrowano dotąd zaledwie 335 publikacji, w których podejmowany był temat big data. Wielkość taka stanowi zaledwie 0,76% wszystkich publikacji o tej tematyce, toteż publikacje te włączone zostały do grupy „Other".

Podsumowując, można zatem stwierdzić, że wykorzystanie technologii big data w naukach społecznych stanowi nurt badań już wyraźnie wyodrębniony i ilościowo znaczący, natomiast w humanistyce zainteresowanie tym modelem badań jest jeszcze niewielkie.



Rys. 4. Rozkład piśmiennictwa dotyczącego big data
według dyscyplin naukowych [wyszukiwanie: 18.11.2018]

W kategoryzacji dyscyplin naukowych stosowanej w bazie Scopus nauka o informacji ulokowana jest w podkategorii Library and Information Sciences, należącej do nauk społecznych. Narzędzia analityczne bazy nie umożliwiają jednak wyodrębnienia podkategorii dziedzinowych w wynikach wyszukiwania, dlatego też bardziej szczegółowe analizy dotyczące problematyki danych masowych w nauce o informacji przeprowadzone zostały na podstawie piśmiennictwa zarejestrowanego w dziedzinowej bazie EBSCO – Library and Information Science and Technology Abstracts (LISTA). Wyniki tych badań omówione są w następnej części artykułu.

3. Big data i analityka danych masowych a problematyka badawcza nauki o informacji

Zagadnienie big data można traktować jako kolejną fazę rozwoju technologii komputerowej i jej zastosowań w różnych dziedzinach nauki i praktyki. Wokół problematyki wykorzystania technologii komputerowej do zapewnienia sprawnego dostępu do utrwalonej informacji i wiedzy w połowie XX w. ukształtowała się współczesna nauka o informacji jako interdyscyplinarny obszar badań o własnym, jednoznacznie wyodrębnionym repertuarze problemów badawczych. Tefko Saracevic, analizując specyfikę tych problemów, opisał naukę o informacji następująco:

> Information science is the science and practice dealing with the effective collection, storage, retrieval, and use of information. It is concerned with recordable information and knowledge, and the technologies and related services that facilitate their management and use. More specifically, information science is the field of professional practice and scientific inquiry addressing the effective communication of information and information objects, particularly knowledge records, among humans in the context of social, organizational, and individual need for and use of information (Saracevic, 2010, 2570).

Z kolei według popularnego *Online Dictionary of Library and Information Science* nauka o informacji to:

> The systematic study and analysis of the sources, development, collection, organization, dissemination, evaluation, use, and management of information in all its forms, including the channels (formal and informal) and technology used in its communication (Reitz, 2014).

Warto tu jeszcze przytoczyć wyróżnienie przez Bruno Jacobfeuerborna dwóch wymiarów nauki o informacji (Jacobfeuerborn, 2013). W pierwszym, wykorzystuje ona interdyscyplinarne podejście do rozwijania podstaw teoretycznych i metodologicznych własnych problemów badawczych, skupionych na społecznym transferze informacji i wiedzy oraz szerokim kontekście jego uwarunkowań. W wymiarze drugim, celem nauki o informacji jest pomaganie badaczom, uczonym, inżynierom, wynalazcom i innym tzw. pracownikom wiedzy w lokalizacji i pozyskaniu informacji i wiedzy niezbędnych w ich pracy. W obu tych wymiarach zjawisko big data powinno być istotnym elementem problematyki badawczej dyscypliny. W pierwszym, może pomóc w znalezieniu rozwiązania problemów związanych z zarządzaniem informacją i wiedzą w środowisku ogromnych i szybko rosnących zbiorów danych, w drugim – może wskazać skuteczne metody obsługi użytkowników w tym nowym środowisku informacyjnym. Odpowiedzią na potrzeby związane z rozwiązaniem tych problemów jest kształtowanie się w ostatnich latach tzw.

data science – nowego nurtu badań w szeroko rozumianej nauce o informacji skupionego na poszukiwaniu i praktycznym zastosowaniu metod derywowania wiedzy (znaczenia i wartości) z wielkich zbiorów danych.

W środowisku wielkich zasobów danych zapisanych w cyfrowym formacie, technologie big data zapewniają wgląd w wiedzę, której nie można byłoby wydobyć tradycyjnymi metodami wyszukiwania informacji. W tym sensie można powiedzieć, że technologie te wspierają procesy transferu wiedzy między ludźmi (jakkolwiek nie tylko między ludźmi[5]), które stanowią główny przedmiot zainteresowań nauki o informacji. A zatem zasadna jest teza, że rozwój technologii big data i jej zastosowań stanowi nowe i niezwykle ważne wyzwanie dla nauki o informacji, pozostając w ścisłej korelacji z jej kluczową problematyką badawczą. Teza ta powinna znaleźć potwierdzenie w rosnącej liczbie badań dotyczących big data w nauce o informacji. Omówiona poniżej analiza piśmiennictwa zarejestrowanego w bazie LISTA jest próbą weryfikacji tej tezy.

3.1. Metoda i próba badawcza

Jak wspomniano wcześniej, zainteresowanie problematyką big data wśród badaczy nauki o informacji zostało zbadane na podstawie piśmiennictwa indeksowanego w międzynarodowej, dziedzinowej bazie LISTA (EBSCO), która należy do najbardziej wyczerpujących źródeł informacji o piśmiennictwie naukowym tej dyscypliny. W bazie tej indeksowanych jest ponad 560 czasopism z zakresu nauki o informacji i bibliotekoznawstwa (NIB) oraz ich dyscyplin pokrewnych, a także wybrane książki i materiały konferencyjne. W LISTA indeksowane jest piśmiennictwo wydawane w ponad 20 językach, jakkolwiek zdecydowaną większość stanowią publikacje w języku angielskim. Zasięg chronologiczny bazy obejmuje okres od połowy lat 60. XX w. Trzeba zaznaczyć, że – ponieważ zakres tematyczny bazy obejmuje nie tylko piśmiennictwo nauki o informacji, a baza nie zapewnia możliwości automatycznego wyodrębniania publikacji reprezentujących poszczególne subdyscypliny objęte indeksowaniem – trudno uzyskane wyniki jednoznacznie interpretować jako dotyczące wąsko rozumianej nauki o informacji. Notabene, na ogół niemożliwe jest wyznaczenie granic pola badawczego nauki o informacji jednoznacznie oddzielających je od problematyki bibliotekoznawstwa, tzw. informatyki stosowanej i wielu innych dziedzin, w których podejmowane są badania nad zjawiskami i procesami informacyjnymi. Stąd wyniki przeprowadzonego wyszukiwania trzeba interpretować w odniesieniu do całego obszaru tematycznego, objętego indeksowaniem w bazie LISTA. Aby jednak ocenić rolę badań dotyczących big data w nauce o informacji w sensie omówionym w poprzedniej

[5] Tradycyjnie przyjmuje się, że jednym z trzech głównych obszarów badań, składających się na intelektualną strukturę nauki o informacji, obok problematyki źródeł informacji i problematyki technologii informacyjnej, jest problematyka użytkowników informacji i użytkowania informacji, która dotychczas była łączona przede wszystkim z badaniem potrzeb i zachowań informacyjnych ludzi (por. Sosińska-Kalata, 2017). Rozwój inteligentnych technologii informacyjnych zmusza jednak do weryfikacji koncepcji użytkownika informacji, włączając do niej również problematykę użytkowania informacji przez np. inteligentne urządzenia, wspierające, a coraz częściej zastępujące człowieka w różnych działaniach związanych z poszukiwaniem i pozyskiwaniem potrzebnej informacji. W polskim piśmiennictwie kwestię konieczności redefinicji pojęcia użytkownika w nauce o informacji omawiał ostatnio Remigiusz Sapa (2018).

części artykułu, przyjrzano się też obecności tej problematyki na łamach czasopism, które dotychczas uznawane były za najbardziej reprezentatywne dla tej dyscypliny[6].

Wyszukiwanie przeprowadzone zostało 15 listopada 2018 r. za pomocą trzech kwerend, które miały zapewnić:

(a) wyodrębnienie najwcześniejszych wystąpień określenia „big data" w piśmiennictwie zarejestrowanym w LISTA;
(b) wyodrębnienie najwcześniejszych wystąpień określenia „big data" w czasopismach naukowych (recenzowanych) indeksowanych w LISTA;
(c) wyodrębnienie zarejestrowanego w LISTA piśmiennictwa naukowego, w którym podejmowano problematykę danych masowych (big data).

Analiza ilościowa piśmiennictwa naukowego, w którym podejmowano problematykę big data, została przeprowadzona według pięciu kryteriów:

(a) data publikacji (rozkład chronologiczny badań);
(b) język publikacji;
(c) czasopisma (koncentracja i rozproszenie publikacji o big data);
(d) afiliacje autorów (rozkład geograficzny badań);
(e) struktura tematyczna.

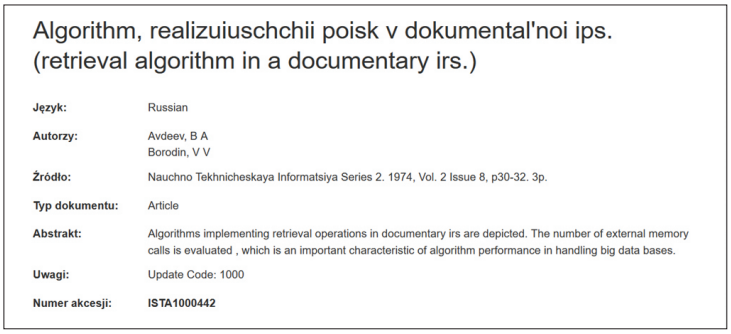
Trzeba zaznaczyć, że omówione poniżej badanie ma charakter sondażowy i przeprowadzone zostało na indeksowanym w bazie LISTA piśmiennictwie, które wyodrębniono za pomocą prostej kwerendy wymagającej jedynie użycia w opisie tematycznym terminu „big data". Piśmiennictwo prezentujące problematykę związaną z badaniem wielkich zbiorów danych, w którego indeksowaniu nie użyto tego terminu, nie zostało zatem uwzględnione w badaniu.

3.2. Najwcześniejsze wystąpienia określenia „big data" w piśmiennictwie zarejestrowanym w bazie LISTA

Ogólny sondaż najwcześniejszych wystąpień problematyki big data w piśmiennictwie zarejestrowanym w bazie LISTA został przeprowadzony na podstawie kwerendy: „big data" (wszystkie pola). Poszukiwano więc wystąpień wyrażenia „big data" w całej zawartości rekordów. W rezultacie uzyskano 1288 rekordów publikacji z okresu 1974–2018.

[6] Za czasopisma najbardziej reprezentatywne („główne", „kanoniczne") dla nauki o informacji uważa się takie czasopisma, których profil tematyczny obejmuje szerokie spektrum problematyki badawczej tej dyscypliny, wokół których skupiają się jej uznani badacze i które należą do najczęściej cytowanych, co znajduje odwzorowanie w wysokim wskaźniku wpływu (IF, SNIP). Na podstawie takich kryteriów do grupy najbardziej reprezentatywnych czasopism nauki o informacji obecnie należałoby zaliczyć: *International Journal of Information Management*, *Information Processing & Management*, *Journal of the Association for Information Science and Technology*, *Journal of Information Science*, *Aslib Journal of Information Management*, *Journal of Documentation, Information Research.* Wielu badaczy do tej grupy dodaje też *Scientometrics, Journal of Informetrics* oraz *Library Hi Tech* i *Library and Information Science Research* jako czasopisma o wysokim IF, które specjalizują się w subdyscyplinach nauki o informacji należących do jej kluczowych nurtów badawczych. Ze względu na fakt, że piśmiennictwo nauki o informacji jest rejestrowane w bazach zwykle obejmujących zakres szerszy niż pole badawcze tej dyscypliny, tego rodzaju podejście bywa stosowane przez badaczy, którzy podejmują próby oceny zjawisk i własności specyficznych dla samej nauki o informacji, np. stan rozwoju dyscypliny, trendy badawcze, front badań, współpraca międzynardowa, poziom interdyscyplinarności itp. (zob. np. White & McCain, 1998; Zhao & Strotmann, 2008; Chang & Huang, 2012; Sosińska-Kalata, 2013).

Najwcześniejsze publikacje, wydane w latach 1974, 1977 i 1981, dotyczą przetwarzania dużych baz danych, a wyszukanie rekordów tych publikacji wynika z wystąpienia w ich abstraktach frazy „big data" w wyrażeniu „big data bases". Tematem pierwszej z tych publikacji – artykułu, który ukazał się w periodyku *Naučno-techničeskaja informacija* – były algorytmy wyszukiwania informacji w dużych bazach dokumentacyjnych (Rys. 5). Jakkolwiek artykuł ten nie dotyczy technologii big data w sensie współczesnym, to warto zwrócić uwagę na to, iż omawiano w nim problemy związane z przetwarzaniem wielkich zbiorów danych oraz na to, iż ukazał się w jednym z głównych czasopism fachowych zajmujących się problematyką informacji naukowej, wydawanych w tamtym czasie w ZSRR.



Algorithm, realizuiuschchii poisk v dokumental'noi ips. (retrieval algorithm in a documentary irs.)

Język:	Russian
Autorzy:	Avdeev, B A Borodin, V V
Źródło:	Nauchno Tekhnicheskaya Informatsiya Series 2. 1974, Vol. 2 Issue 8, p30-32. 3p.
Typ dokumentu:	Article
Abstrakt:	Algorithms implementing retrieval operations in documentary irs are depicted. The number of external memory calls is evaluated , which is an important characteristic of algorithm performance in handling big data bases.
Uwagi:	Update Code: 1000
Numer akcesji:	**ISTA1000442**

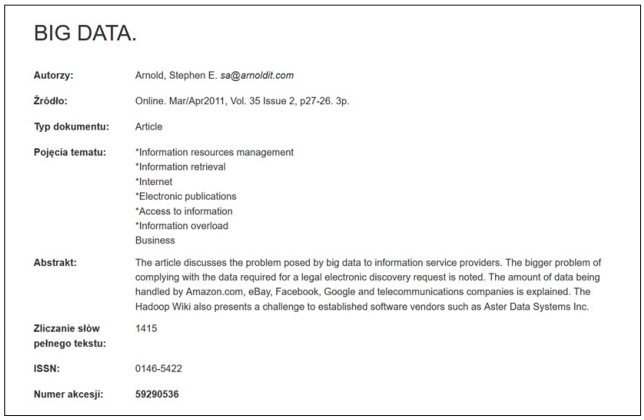
Rys. 5. Pierwsze zarejestrowane w bazie LISTA użycie wyrażenia „big data" w abstrakcie artykułu

W 1983 r. w *Lecture Notes on Computer Science* ukazał się pierwszy artykuł, spośród zarejestrowanych w bazie LISTA, w którym użyte zostało określenie „big data" na oznaczenie wielkich zbiorów danych. Również ten artykuł dotyczył tworzenia nowych algorytmów przeszukiwania tego typu zbiorów danych (Rys. 6).

Treatment of big values in an applicative language HFP. Translation from by-value access to by-update access

Autorzy:	Katayama, T[1]
Źródło:	Lecture Notes on Computer Science. Vol. 147. 1983.
Typ dokumentu:	Book Chapter
Pojęcia tematu:	COMPUTERS ENGINEERING LANGUAGE & languages
Słowa kluczowe podane przez autora:	Access methods
Abstrakt:	This paper proposes a method of treating big data by converting by-value access to by-update access, which is used in the implementation of an applicative language HFP. HFP is an applicative language which admits hierarchical and applicative programming and is based on the attribute grammar of Knuth. It also has a close relationship to Prolog. The author first introduces HFP and discusses its implementation which solves the big data problem by using a simple file processing program Book Published by Springer-Verlag, Germany, 1983
Uwagi:	Place of Publication: Germany Publisher: Springer-Verlag Update Code: 1900
Afiliacje autora:	[1]Tokyo Inst. of Technology Tokyo, Japan
Numer akcesji:	**ISTA1904357**

Rys. 6. Pierwszy artykuł, zarejestrowany w bazie LISTA, w którym wyrażenie „big data" zostało użyte na oznaczenie wielkich zbiorów danych



BIG DATA.

Autorzy: Arnold, Stephen E. *sa@arnoldit.com*

Źródło: Online. Mar/Apr2011, Vol. 35 Issue 2, p27-26. 3p.

Typ dokumentu: Article

Pojęcia tematu: *Information resources management
*Information retrieval
*Internet
*Electronic publications
*Access to information
*Information overload
Business

Abstrakt: The article discusses the problem posed by big data to information service providers. The bigger problem of complying with the data required for a legal electronic discovery request is noted. The amount of data being handled by Amazon.com, eBay, Facebook, Google and telecommunications companies is explained. The Hadoop Wiki also presents a challenge to established software vendors such as Aster Data Systems Inc.

Zliczanie słów pełnego tekstu: 1415

ISSN: 0146-5422

Numer akcesji: **59290536**

Rys. 7. Pierwszy artykuł, zarejestrowany w bazie LISTA, dotyczący problematyki big data w kontekście wykorzystywania danych generowanych przez użytkowników Internetu

Pierwsze publikacje, które dotyczą problemu danych masowych, zostały wydane w 2010 r. Są to krótkie komunikaty z konferencji poświęconych nowym metodom informatycznym oraz recenzja książki Davida Bolliera *The Promise and Peril of Big Data.*

W 2011 r. ukazał się natomiast pierwszy artykuł o problematyce big data w kontekście wykorzystywania nieustrukturyzowanych i różnorodnych danych generowanych przez użytkowników Internetu (serwisów e-commerce i mediów społecznościowych). Artykuł ten został wydany w nierecenzowanym magazynie *Online* (Rys. 7).

3.3. Najwcześniejsze wystąpienia określenia „big data" w czasopismach naukowych indeksowanych w bazie LISTA

Do śledzenia najwcześniejszych wystąpień określenia „big data" w czasopismach naukowych zostało wykorzystane ograniczenie wyników poprzedniej kwerendy za pomocą kryterium „czasopisma naukowe (recenzowane)". Wyodrębniono w ten sposób zbiór 808 rekordów, w tym 741 rekordów artykułów naukowych i 67 rekordów artykułów recenzyjnych. Artykuły te ukazały się w latach 2008–2018.

Pierwszym recenzowanym artykułem naukowym, zarejestrowanym w bazie LISTA, w którym stwierdzono wystąpienie określenia „big data" w abstrakcie, jest artykuł kubańskich badaczy wydany w języku hiszpańskim w 2008 r. w czasopiśmie *Ciencias de la Información*. Tematem tego artykułu jest badanie z zakresu patentometrii, tj. analiza danych o kubańskich patentach zarejestrowanych w amerykańskich wielkich bazach patentowych, której celem jest wskazanie najbardziej innowacyjnych kubańskich technologii, ośrodków badawczych i badaczy. A zatem, podobnie jak w przypadku piśmiennictwa nierecenzowanego, również pierwsze użycia określenia „big data" w artykułach recenzowanych wiążą się z analizami wielkich baz danych.

Pierwszym zarejestrowanym w LISTA artykułem badawczym, w którym termin „big data" pojawia się wśród słów kluczowych, identyfikujących główne pojęcia omawianego tematu, jest artykuł wydany w 2011 r. w czasopiśmie *Journal of the American Medical Informatics Association,* dotyczący wykorzystania metod data-mining w symulacjach komputerowych (Rys.8).

The TOKEn project: knowledge synthesis for in silico science.

Autorzy: Payne, Philip R. O.[1] *philip.payne@osumc.edu*
Borlawsky, Tara B.[1]
Lele, Omkar[1]
James, Stephen[1]
Greaves, Andrew W.[2]

Źródło: Journal of the American Medical Informatics Association. 2011 Supplement, Vol. 18, pi125-i131. 7p.

Typ dokumentu: Article

Pojęcia tematu: *Data mining
*Big data
*Database research
*Acquisition of data
Semantics research

Abstrakt: Objective The conduct of investigational studies that involve large-scale data sets presents significant challenges related to the discovery and testing of novel hypotheses capable of supporting in silico discovery science. The use of what are known as Conceptual Knowledge Discovery in Databases (CKDD) methods provides a potential means of scaling hypothesis discovery and testing approaches for large data sets. Such methods enable the high-throughput generation and evaluation of knowledge-anchored relationships between complexes of variables found in targeted data sets. Methods The authors have conducted a multipart model formulation and validation process, focusing on the development of a methodological and technical approach to using CKDD to support hypothesis discovery for in silico science. The model the authors have developed is known as the Translational Ontology-anchored Knowledge Discovery Engine (TOKEn). This model utilizes a specific CKDD approach known as Constructive Induction to identify and prioritize potential hypotheses related to the meaningful semantic relationships between variables found in large-scale and heterogeneous biomedical data sets. Results The authors have verified and validated TOKEn in the context of a translational research data repository maintained by the NCI-funded Chronic Lymphocytic Leukemia Research Consortium. Such studies have shown that TOKEn is: (1) computationally tractable, and (2) able to generate valid and potentially useful hypotheses concerning relationships between phenotypic and biomolecular variables in that data collection. Conclusions The TOKEn model represents a potentially useful and systematic approach to knowledge synthesis for in silico discovery science in the context of large-scale and multidimensional research data sets. [ABSTRACT FROM AUTHOR]

Afiliacje autora: [1]Department of Biomedical Informatics, Ohio State University, Columbus, Ohio, USA
[2]CLL Research Consortium, Moores UCSD Cancer Center, La Jolla, California, USA

ISSN: 1067-5027

DOI: 10.1136/amiajnl-2011-000434

Numer akcesji: 103564366

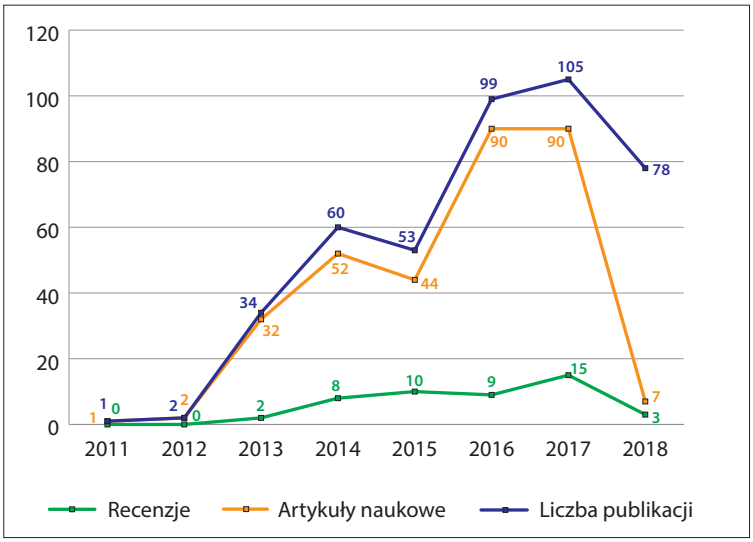
Rys. 8. Pierwszy zarejestrowany w bazie LISTA artykuł badawczy, w którego rekordzie termin „big data" występuje w polu pojęć tematu

3.4. Piśmiennictwo nauki o informacji, bibliotekoznawstwa i nauk pokrewnych, w którym podejmowano problematykę big data

W celu wyodrębnienia z bazy LISTA piśmiennictwa naukowego, którego przedmiotem są badania dotyczące big data, została użyta kwerenda, w której poszukiwane wystąpienia terminu „big data" ograniczono do pola pojęć tematu (SU) oraz typ dokumentu ograniczono do kategorii „czasopisma naukowe (recenzowane)". Na podstawie tej kwerendy otrzymano zbiór 427 rekordów, w tym 381 rekordów artykułów naukowych i 47 rekordów recenzji w czasopismach naukowych. Publikacje te ukazały się w latach 2011–2018, a więc w ciągu ostatnich ośmiu lat. Najwcześniejszym artykułem zaindeksowanym terminem „big data" jest wspomniany w poprzedniej części artykuł z zakresu informatyki medycznej o zastosowaniu technik data-mining w symulacjach komputerowych.

3.4.1. Rozkład chronologiczny

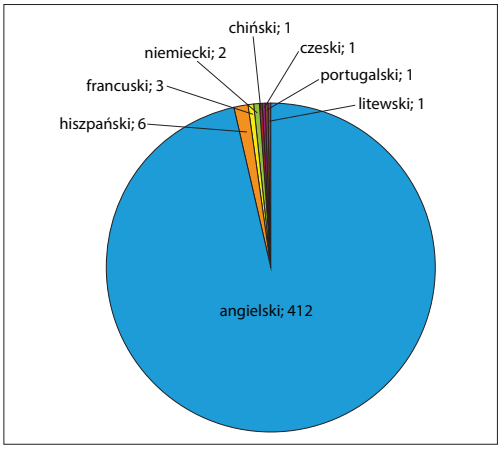
Chronologiczny rozkład publikacji na temat big data w czasopismach naukowych indeksowanych w bazie LISTA ukazuje systematyczny wzrost zainteresowania tą problematyką w ostatnich ośmiu latach (Rys. 9). Dane te pokazują też, iż, podobnie jak w innych dyscyplinach naukowych, również w nauce o informacji szybkie zwiększanie się liczby badań prowadzonych w tym zakresie następuje od 2011 r. W listopadzie 2018 r., kiedy przeprowadzane było wyszukiwanie, dane za rok 2018 były niepełne, stąd nie można brać ich pod uwagę w ocenie trendu.



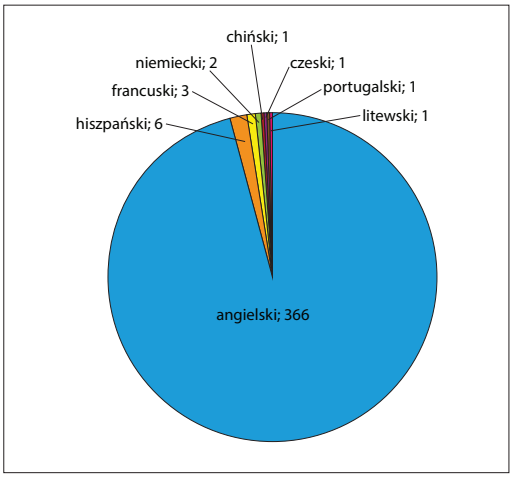
Rys. 9. Rozkład chronologiczny publikacji nt. big data w czasopismach naukowych indeksowanych w bazie LISTA

3.4.2. Języki publikacji

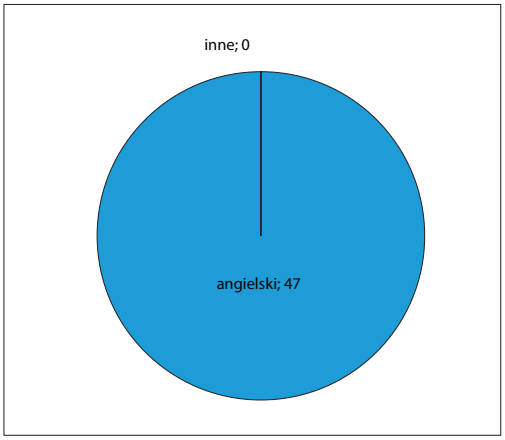
Analiza wyodrębnionego piśmiennictwa według języków publikacji (Rys. 10, 11 i 12) nie wnosi wiele do poznania specyfiki literatury dotyczącej tematyki big data w nauce o informacji i dziedzinach pokrewnych. Przeprowadzono ją, zakładając, że może ujawnić szczególne zainteresowanie tą problematyką w pewnych kręgach językowych, z wyłączeniem języka angielskiego, który jest podstawowym językiem komunikacji naukowej w obiegu międzynarodowym. Wyniki analizy nie pozwalają jednak na formułowanie takich wniosków.



Rys. 10. Rozkład wg języków publikacji wszystkich publikacji naukowych o big data, zarejestrowanych w bazie LISTA



Rys. 11. Rozkład wg języków publikacji artykułów naukowych o big data, zarejestrowanych w bazie LISTA



Rys. 12. Rozkład wg języków publikacji recenzji książek o big data, wydanych ma łamach czasopism naukowych indeksowanych w bazie LISTA

Piśmiennictwo naukowe o big data, które zarejestrowano w bazie LISTA, wydano w ośmiu językach. Jak można było spodziewać się, przytłaczająca większość publikacji (97%) ukazała się w języku angielskim. W pozostałych językach wydano w ciągu ostatnich ośmiu lat od sześciu do jednego artykułu: sześć w języku hiszpańskim, trzy w języku francuskim, dwa w języku niemieckim oraz po jednym w językach: chińskim, czeskim, litewskim i portugalskim. Zwraca uwagę relatywnie większa liczba publikacji w języku hiszpańskim niż w językach pozostałych, jednak z reguły wielkości te są zbyt małe, aby można było mówić o szczególnym zainteresowaniu badaniami big data badaczy, którzy publikują w którymkolwiek języku innym niż język angielski. Także sam zestaw języków, w których opublikowano dotąd po jednym artykule, trzeba traktować jako dość przypadkowy. Uzyskane

dane demonstrują przede wszystkim dominację języka angielskiego zarówno w polityce indeksowania bazy LISTA, jak i ogólnie we współczesnej komunikacji naukowej.

3.4.3. Czasopisma nauki o informacji, w których ukazały się publikacje o big data

Więcej interesujących informacji dostarcza analiza czasopism, w których ukazały się artykuły i recenzje książek o tematyce big data. Publikacje te wydano na łamach 113 czasopism, czyli w około 20% źródeł indeksowanych w bazie LISTA. Piśmiennictwo to jest więc znacznie rozproszone. Ponad połowa (51,8%) artykułów i recenzji dotyczących zagadnień big data ukazała się jednak w zaledwie 13 czasopismach (Rys. 13). Można zatem stwierdzić, że wśród źródeł indeksowanych w bazie LISTA istnieje względnie nieduża grupa czasopism wyraźnie bardziej zainteresowana zagadnieniami big data niż pozostałe źródła. Biorąc pod uwagę liczbę opublikowanych artykułów, w grupie tej wyróżnić można trzy podgrupy:

(1) czasopisma, w których ukazało się ponad 20 artykułów; podgrupę tę tworzą cztery periodyki (*Scientometrics*, *Information Systems*, *Journal of the American Medical Informatics Association*, *Choice: Current Reviews for Academic Libraries*);
(2) czasopisma, w których ukazało się od 14 do 20 artykułów; do tej podgrupy należą trzy periodyki (*First Monday*, *Information Journal of Information Management*, *Information, Communication and Society*);
(3) czasopisma, w których ukazało się od 8 do 10 artykułów; podgrupę tę tworzy sześć periodyków (*Information Services & Use*, *Library Hi Tech*, *El Profesional de la Informacion*, *Information Processing and Management*, *Journal of Medical Internet Research*, *Information Polity: The International Journal of Government & Democracy in the Information Age*).

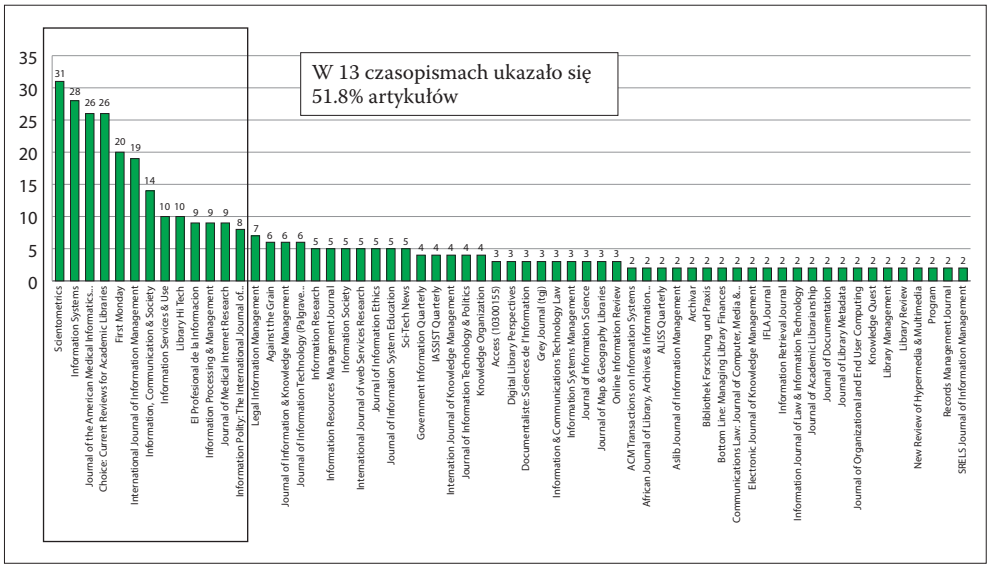
W pierwszej grupie znajdują się trzy periodyki, które zajmują się specjalistycznymi obszarami szeroko rozumianej nauki o informacji (naukometrią, zastosowaniami technologii inteligentnych, informatyką medyczną) oraz czasopismo przeglądowe, wydawane przez Association of College and Research Libraries, publikujące pisane przez naukowców recenzje książek naukowych oraz źródeł internetowych zawierających treści naukowe. Chociaż w bazie LISTA magazyn *Choice* został skategoryzowany jako recenzowane czasopismo naukowe, nie jest to czasopismo publikujące oryginalne artykuły badawcze.

Wśród trzech czasopism tworzących drugą grupę znajdują się czasopisma naukowe, zajmujące się problematyką szeroko rozumianej nauki o informacji, w tym dwa, które mają charakter interdyscyplinarnych czasopism zajmujących się problematyką komunikacji społecznej i współczesnych mediów (*Information, Communication and Society* oraz *First Monday*).

W trzeciej grupie znajdują się cztery czasopisma, których profil obejmuje szeroki zakres problematyki badawczej nauki o informacji (*Information Services & Use*, *Library Hi Tech*, *El Profesional de la Informacion*, *Information Processing and Management*) oraz dwa poświęcone specjalistycznym zagadnieniom nauki o informacji (*Journal of Medical Internet Research*, *Information Polity: The International Journal of Government & Democracy in the Information Age*).

Dane te pozwalają wysnuć wniosek, że w nauce o informacji badaniami big data interesują się w największym stopniu czasopisma, których profil ukierunkowany jest na problematykę zasobów i usług informacyjnych w obszarach specjalistycznych. Nieco rzadziej problematyka ta dotychczas pojawiała się na łamach czasopism nauki o informacji o profilu ogólnym, obejmującym szeroki repertuar problemów badawczych nauki

o informacji. Interesujące jest przy tym to, że wśród czasopism publikujących artykuły dotyczące badań big data jest niewiele takich, które należą do tzw. kanonicznych czy też głównych czasopism nauki o informacji – uznawanych za najważniejsze i najbardziej reprezentatywne dla tej dyscypliny.



Rys. 13. Ilościowy rozkład artykułów dotyczących big data w czasopismach naukowych z zakresu nauki o informacji i dyscyplin pokrewnych, w których ukazały się co najmniej dwa artykuły o tej tematyce

W tabeli 1 zostały zestawione tytuły czasopism, w których w latach 2011–2018 ukazało się co najmniej pięć artykułów poświęconych big data. Obok tytułów czasopism umieszczono ich aktualny[7] *impact factor* (IF), Source Normalized Impact per Paper (SNIP)[8] oraz liczbę artykułów zaindeksowanych w bazie LISTA do 18 listopada 2018 r. Zestawienie pokazuje, że pierwsze trzy czasopisma, w których w badanym okresie ukazało się najwięcej artykułów o analizowanej tu problematyce, należą do wysoko cytowanych periodyków naukowych, rejestrowanych zarówno w Web of Science, jak i w Scopus. Według danych *Journal Citation Reports*, w kategorii „Information Science & Library Science" *Scientometrics* zajmuje 25. pozycję w rankingu czasopism o największym oddziaływaniu, *Information Systems* pozycję szóstą, a *Journal of the American Medical Informatics Association* pozycję piątą. Czasopismo *International Journal of Information Management,* którego IF ma największą wartość wśród czasopism wyodrębnionych z bazy LISTA, w JCR w 2017 r. zostało sklasyfikowane na trzeciej pozycji wśród czasopism należących do kategorii „Information Science & Library Science".

[7] Na podstawie *Journal Citation Reports* bazy Web of Science – raport za 2017 r.
[8] Na podstawie danych o indeksowanych źródłach bazy Scopus za 2017 r.

Tab. 1. Czasopisma indeksowane w bazie LISTA, w których ukazało się co najmniej pięć artykułów dotyczących problematyki big data

L.p.	Tytuł czasopisma	IF (2017)	SNIP (2017)	Liczba artykułów
1	**Scientometrics**	**2.173**	**1.378**	**31**
2	Information Systems	4.267	2.251	28
3	Journal of the American Medical Informatics Association	4.270	2.262	26
4	Choice: Current Reviews for Academic Libraries	–	–	26
5	First Monday	–	0.771	20
6	**International Journal of Information Management**	**4.516**	**2.824**	**19**
7	Information, Communication & Society	–	1.989	14
8	Information Services & Use	–	0.497	10
9	**Library Hi Tech**	**0.759**	**0.722**	**10**
10	El Profesional de la Informacion	–	1.130	9
11	**Information Processing & Management**	**3.444**	**2.66**	**9**
12	Journal of Medical Internet Research	–	1.815	9
13	Information Polity: The International Journal of Government & Democracy in the Information Age	–	0.935	8
14	Legal Information Management	–	–	7
15	Against the Grain	–	–	6
16	Journal of Information & Knowledge Management	–	0.56	6
17	Journal of Information Technology (Palgrave Macmillan)	4.535	2.638	6
18	**Information Research**	**0.762**	**0.815**	**5**
19	Information Resources Management Journal	–	0.209	5
20	Information Society	1.889	1.225	5
21	International Journal of Web Services Research	–	0.387	5
22	Journal of Information Ethics	–	0.201	5
23	Journal of Information Systems Education	–	0.685	5
24	Sci-Tech News	–	–	5

Tabela 1 pokazuje też, iż więcej niż połowa czasopism, w których ukazały się artykuły o badaniach dotyczących big data w kontekście nauki o informacji i jej dziedzin pokrewnych, to periodyki, które nie są rejestrowane w Web of Science, a zatem o poziomie wpływu niższym niż wymagany od czasopism objętych tą bazą. W przypadku indeksowania w bazie Scopus, tylko cztery spośród tych czasopism nie są nim objęte.

W tabeli 1 czcionką półgrubą zostały zaznaczone czasopisma, które należą do tzw. kanonicznych czasopism nauki o informacji. Wśród periodyków, w których ukazało się co najmniej pięć artykułów o big data jest ich zaledwie pięć. Pełniejszy obraz obecności tej problematyki w czasopismach, które można uznać za najważniejsze we współczesnej nauce o informacji, prezentuje tabela 2. W okresie objętym badaniem w czasopismach tych opublikowano łącznie 82 artykuły dotyczące problematyki big data, co stanowi nieco więcej niż jedną piątą wszystkich publikacji zaindeksowanych terminem „big data" w bazie LISTA.

W tej grupie najwięcej, bo blisko 38% publikacji, ukazało się w *Scientometrics*, podobnie zresztą jak w całej badanej próbie (ponad 8%). Ilościowe badania informacji o nauce i piśmiennictwie naukowym wyraźnie stanowią nurt badań, w którym technologie i metody big data znajdują najczęstsze zastosowanie.

Tab. 2. „Kanoniczne" czasopisma nauki o informacji, w których ukazały się artykuły dotyczące problematyki big data

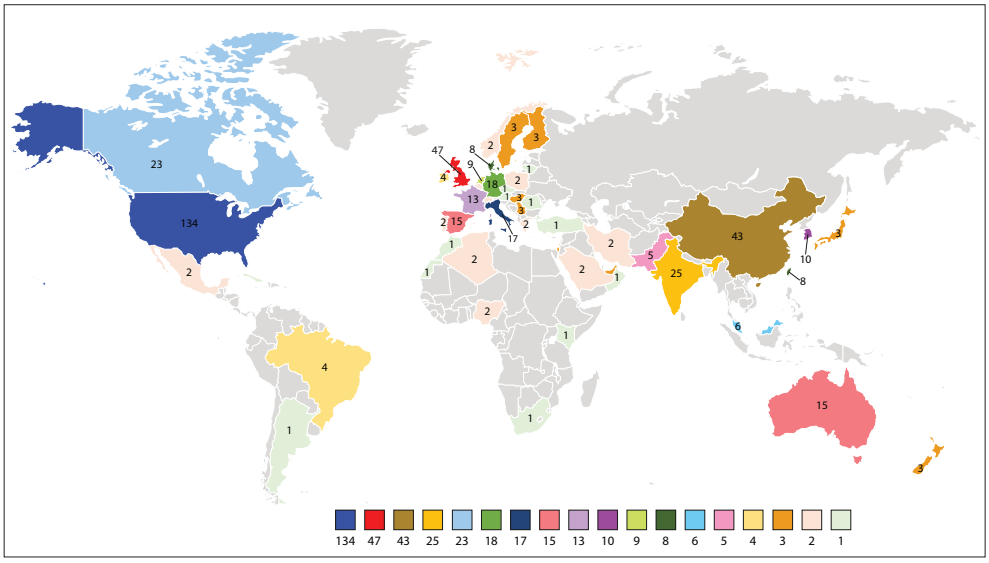
Lp.	Tytuł czasopisma	IF (2017)	SNIP (2017)	Liczba artykułów
1	**Scientometrics**	**2.173**	**1.378**	**31**
2	**International Journal of Information Management**	**4.516**	**2.824**	**19**
3	**Library Hi Tech**	**0.759**	**0.722**	**10**
4	**Information Processing & Management**	**3.444**	**2.66**	**9**
5	**Information Research**	**0.762**	**0.815**	**5**
6	Journal of Information Science	1.939		3
7	Aslib Journal of Information Management	1.461		2
8	Journal of Documentation	1.157		2
9	Journal of the Association for Information Science & Technology	2.835		1
Razem artykułów				82 (21.85%)

Zaskakujące może wydawać się to, że w ciągu ostatnich ośmiu lat zaledwie jeden artykuł o tematyce big data ukazał się w JASIST (*Journal of the Association for Informaion Science and Technology*, do 2013 r. *Journal of the American Society for Information Science and Technology*) – czasopiśmie często uznawanym za najważniejsze źródło dla nauki o informacji i najbardziej reprezentatywne dla jej pola badawczego. Równie zastanawiające jest to, że w *Journal of Documentation*, najstarszym europejskim czasopiśmie poświęconym tej dyscyplinie i także uznawanym za należące do najbardziej dla niej reprezentatywnych, dotychczas ukazały się tylko dwa artykuły dotyczące big data. Trzeba jednak przypomnieć, że wyniki przeprowadzonej analizy piśmiennictwa indeksowanego w bazie LISTA wskazało, iż liczba publikacji o tematyce big data w czasopismach o szerokim profilu, obejmującym różne obszary badań nauki o informacji, jest wyraźnie mniejsza niż w czasopismach o profilu bardziej specjalistycznym. Wyjątek stanowi *International Journal of Information Management*, którego profil określony jest względnie szeroko, ale w którym dominuje jednak problematyka zarządzania informacją w organizacjach, a więc dotycząca dziedziny, w której zastosowania technologii big data należą do najczęstszych.

3.4.4. Geograficzny rozkład publikacji naukowych na temat big data w nauce o informacji

Na podstawie analizy afiliacji autorów publikacji naukowych wyodrębnionych z bazy LISTA został ustalony rozkład geograficzny tego piśmiennictwa, który interpretować można też jako geograficzny rozkład ośrodków badawczych, zajmujących się zagadnieniami big data w obszarze problemowym nauki o informacji (Rys. 14 i 15). Rozkład ten jednoznacznie wskazuje na zdecydowaną dominację amerykańskich ośrodków badawczych: w Stanach Zjednoczonych znajdują się ośrodki badawcze, przy których afiliowana jest blisko jedna

trzecia (29%) autorów artykułów o tej tematyce. Do grupy krajów, w których znajduje się najwięcej ośrodków afiliujących badaczy zajmujących się big data w nauce o informacji należą: Wielka Brytania (10%) i Chiny (9%). Na Stany Zjednoczone, Wielką Brytanię i Chiny przypada łącznie prawie połowa badań o tej tematyce, które omówione zostały w piśmiennictwie indeksowanym w LISTA. W Indiach i w Kanadzie zlokalizowanych jest po 5% ośrodków, przy których afiliowani są autorzy badanego piśmiennictwa, w Niemczech i we Włoszech po 4%, a na Australię, Hiszpanię i Francję przypada po 3% ośrodków.

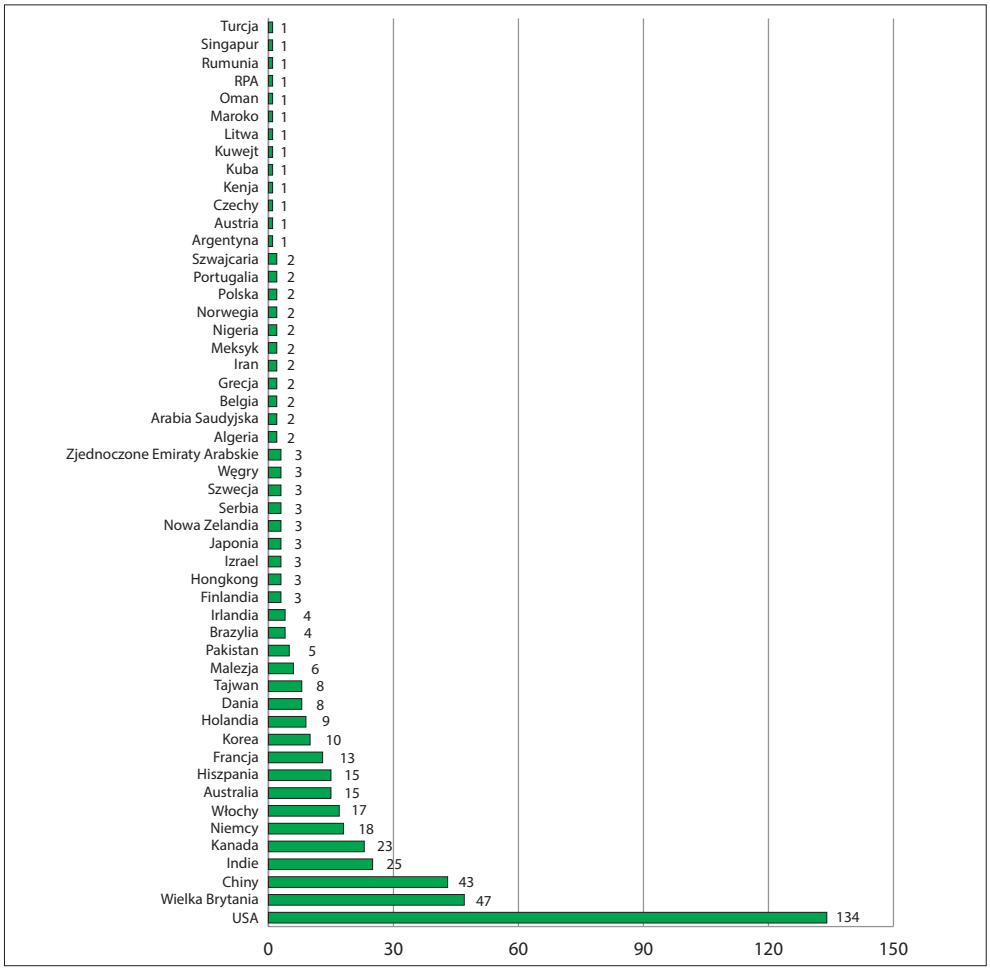


Rys. 14. Geograficzny rozkład publikacji naukowych dotyczących big data, zarejestrowanych w bazie LISTA

Nie jest zaskoczeniem fakt, że najbardziej aktywni w problematyce dotyczącej big data i zastosowań tej technologii w obszarze problemowym nauki o informacji są badacze związani z ośrodkami naukowymi w Stanach Zjednoczonych. Powszechnie znana jest supremacja tego kraju na polu badań nowych technologii i ich zastosowań, i tam nauka o informacji niemal od początku rozwija się najbardziej prężnie. Nie można zapominać też, że to właśnie amerykańskie firmy (Microsoft, Google, Amazon, Apple) mają najlepszy dostęp do globalnych big data. Również wysoka intensywność badań dotyczących problematyki informacyjnej zarówno w wymiarze technologicznym, jak i społecznym w Wielkiej Brytanii znajduje potwierdzenie w znaczącym zainteresowaniu tamtejszych ośrodków badawczych tą nową technologią i jej aplikacjami. Ciekawym zjawiskiem jest duża liczba chińskich ośrodków naukowych zaangażowanych w nurt badań dotyczących big data w nauce o informacji. Coraz bardziej intensywna eksploracja zaawansowanych technologii przez chińskich badaczy jest widoczna także w wielu innych dziedzinach. Warto zauważyć, że szybki rozwój takich firm jak Alibaba, Baidu czy Hauwei zapewnia Chinom coraz lepszy dostęp do big data.

Porównanie rozkładu geograficznego afiliacji autorów publikacji dotyczących big data w nauce o informacji z omówionym wcześniej analogicznym rozkładem piśmiennictwa

z różnych dziedzin zarejestrowanego w bazie Scopus (Rys. 3) pozwala stwierdzić, że chińska nauka jest obecnie wyraźnym liderem w badaniach nad big data na świecie, jednak w nauce o informacji Stany Zjednoczone i Wielka Brytania nadal utrzymują nad nią przewagę. Poza tym oba rozkłady są dość podobne. Interesującą kwestią jest też zaobserwowana, w piśmiennictwie o tematyce big data zarejestrowanym w bazie LISTA, stosunkowo duża liczba afiliowanych w ośrodkach amerykańskich autorów, których nazwiska sugerują ich chińskie pochodzenie. Wyciągnięcie wniosków z tej obserwacji wymaga jednak dokładniejszych badań.



Rys. 15. Ilościowy rozkład publikacji naukowych dotyczących zagadnień big data według państw, w których znajdują się ośrodki badawcze afiliujące ich autorów

Uwagę bez wątpienia zwraca nieobecność Rosji w geograficznym rozkładzie piśmiennictwa dotyczącego big data w nauce o informacji. Można przypuszczać, że pewien wpływ na taki stan rzeczy ma polityka indeksowania bazy LISTA, preferująca publikacje w języku

angielskim, jednakże w bazie tej zaindeksowano łącznie 756 publikacji wydanych w języku rosyjskim, z czego 684 to teksty opublikowane w okresie 2011–2018. Z dużym prawdopodobieństwem problem dotyczy więc zainteresowania zjawiskiem big data rosyjskich badaczy zajmujących się nauką o informacji. W interdyscyplinarnej bazie Scopus wśród 44 052 publikacji zaindeksowanych słowem kluczowym „big data" 590 ma autorów afiliowanych w rosyjskich ośrodkach badawczych, z czego 539 opublikowano w języku angielskim, a tylko 51 w języku rosyjskim. W bazie Scopus, w zestawieniu piśmiennictwa o big data według krajów afiliacji autorów, Federacja Rosyjska zajmuje 14 pozycję.

Wśród piśmiennictwa o big data zaindeksowanego w LISTA znalazły się dwie publikacje autorów z polskich ośrodków badawczych, obie wydane w języku angielskim. W bazie Scopus zarejestrowano 358 publikacji o tej tematyce autorów afiliowanych w polskich instytucjach naukowych, co w zestawieniu piśmiennictwa według krajów afiliacji lokuje Polskę na 27. miejscu.

3.4.5. Tematyka publikacji dotyczących problematyki big data w nauce o informacji

Próbę ustalenia struktury tematycznej piśmiennictwa o zagadnieniach big data zarejestrowanego w bazie LISTA podjęto na podstawie analizy słów kluczowych występujących w polu tematu. Analiza ta została przeprowadzona w dwóch wariantach: na podstawie danych o całym zbiorze 381 artykułów naukowych dotyczących problematyki big data, które zarejestrowano w bazie LISTA oraz na podstawie danych o zbiorze zarejestrowanych w LISTA 82 artykułów na ten temat, opublikowanych w dziewięciu czasopismach uznawanych za najbardziej reprezentatywne dla badań prowadzonych w nauce o informacji.

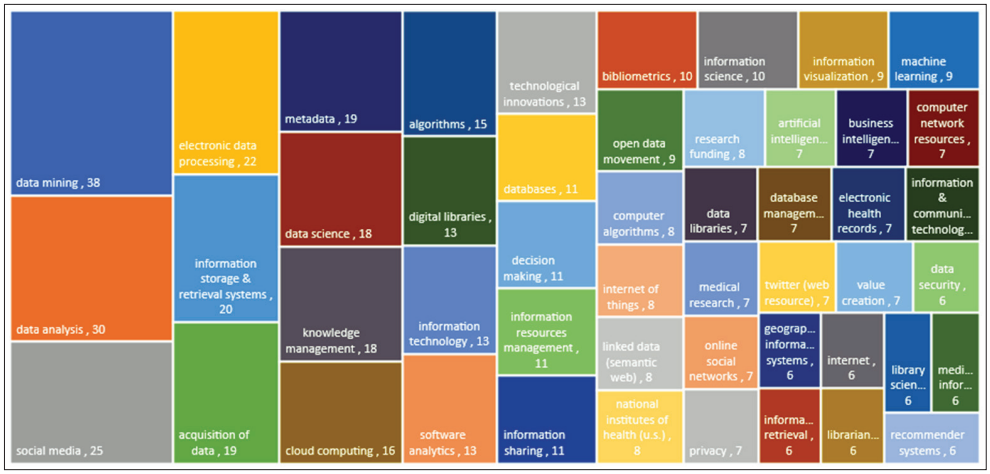
Strukturę tematyczną całego zbioru artykułów naukowych dotyczących big data, które zarejestrowano w bazie LISTA, prezentuje rysunek 16. Ze względu na zapewnienie czytelności wykresu, prezentacja ilościowego rozkładu słów kluczowych została ograniczona do terminów użytych w indeksowaniu co najmniej sześciu publikacji. W analizie nie uwzględniono wystąpień terminu *big data*, którym zaindeksowane były wszystkie artykuły w badanym zbiorze.

Uzyskany rozkład słów kluczowych przede wszystkim uwidacznia dużą różnorodność omawianych zagadnień szczegółowych. Ze zrozumiałych względów wśród terminów użytych w indeksowaniu badanego piśmiennictwa najczęściej występują wyrażenia dotyczące zagadnień informatycznych (*data mining*, *data analysis)* oraz termin *social media*, odnoszący się do najczęściej wykorzystywanego źródła nich wykorzystywanych w analizach big data.

Jeśli do liczby wystąpień terminu *social media* (25) dodamy liczbę wystąpień nazwy *Twitter (web resources)* (7) oraz termin *online social networks* (7), to otrzymamy wartość 39, wyższą nawet od liczby wystąpień terminu *data mining* (38), jak i *data analysis* (30). Możemy zatem powiedzieć, że pojęcia eksploracji danych, analizy danych oraz sieci społecznych w przestrzeni cyfrowej wskazują podstawową ramę konceptualną tematyki badań dotyczących big data w nauce o informacji i jej dziedzinach pokrewnych.

Terminy o nieco mniejszej frekwencji ukazują specyficzne dla nauki o informacji zagadnienia badań szczegółowych, związanych ze zjawiskiem big data. W tej grupie także wskazać można terminy dotyczące zagadnień informatycznych i terminy odnoszące się do obszarów zastosowań analiz big data oraz ich społecznych aspektów. Wśród zagadnień informatycznych, które wydają się w największym stopniu, interesować badaczy szeroko rozumianych nauk informacyjnych występują: *electronic data processing, information*

storage & retrieval systems, acquisition of data, metadata, data science, cloud computing, algorithms, data science, information technology, software analysis, technological innovations, information visualization, machine learning, artificial intelligence, database management, linked data (semantic web). Warto zauważyć, że wśród tych zagadnień widoczne są tematy typowe dla badań nauki o informacji (np. gromadzenie, przechowywanie, wyszukiwanie informacji, problematyka metadanych, zarządzanie bazami danych) oraz tematy związane z zastosowaniem innowacyjnych technologii (sztuczna inteligencja, maszynowe uczenie się, przetwarzanie w chmurze, wizualizacja informacji). Z kolei wśród terminów odnoszących się do obszarów zastosowań analiz i technologii big data oraz społecznych aspektów tego typu badań względnie wysoką częstotliwość w badanym zbiorze mają wyrażenia *knowledge management, digital libraries, decision making, information resources management, information sharing, bibliometrics, information science, open data movement, national institutes of health (US), medical research* i *ecelctronic health records.* Warto zwrócić uwagę na dość znaczny udział wśród analizowanych słów kluczowych terminów związanych z ochroną prywatności i bezpieczeństwem danych.



Rys. 16. Struktura tematyczna artykułów naukowych dotyczących zagadnienia big data, zarejestrowanych w bazie LISTA

Celem analizy słów kluczowych użytych do zaindeksowania 82 artykułów opublikowanych w dziewięciu „kanonicznych" czasopismach nauki o informacji (zob. Tab. 2) jest porównanie struktury tematycznej piśmiennictwa badawczego reprezentującego cały obszar nauki o informacji, bibliotekoznawstwa oraz dyscyplin pokrewnych ze strukturą tematyczną publikacji na ten temat, które uznać można za piśmiennictwo najbardziej reprezentatywne dla nauki o informacji. Jak widać na rysunku 17, rozkład słów kluczowych charakteryzujących tematykę artykułów z czasopism „kanonicznych" nieco różni się od rozkładu prezentującego strukturę tematyczną całego zbioru piśmiennictwa o big data wyodrębnionego z bazy LISTA. W głównych czasopismach nauki o informacji problematyka big data podejmowana była przede wszystkim w kontekście bibliometrii (*bibliometrics*) i bibliotek cyfrowych (*digital libraries*). Można stwierdzić, że temat big data

w tej grupie czasopism podejmowany był najczęściej w kontekście zagadnień ilościowych badań informacji (po pięć wystąpień terminów: *altmetrics, citation analysis*, *scientometrics,* trzy wystąpienia *impact factor* (*citation analysis*) oraz dwa wystąpienia *h-index (citation analysis)*). Niewątpliwie duży wpływ na to ma umieszczenie czasopisma *Scientometrics* w grupie kanonicznych periodyków nauki o informacji. Obecność terminów *data mining* i *social media* wśród wyrażeń najczęściej występujących w polu tematu także w tej grupie artykułów uzasadnić należy analogicznie, jak w przypadku całego zbioru publikacji o tematyce big data wyodrębnionego z bazy LISTA. Terminy te wskazują zatem najczęstszy aspekt technologiczny badań big data i najczęstsze źródło danych poddawanych analizie. Warto natomiast zauważyć, że na czasopisma „kanoniczne" przypada większość spośród publikacji, w których problematykę big data podejmowano w kontekście bibliotek cyfrowych (9 spośród 13). W czasopismach „kanonicznych" relatywnie często omawiano też zagadnienia big data w kontekście systemów wyszukiwania informacji – na czasopisma te przypada 35% wszystkich publikacji na ten temat zarejestrowanych w bazie LISTA. Z kolei w kontekście metadanych o badaniach big data w „kanonicznych" czasopismach nauki o informacji pisano dotąd w pięciu artykułach, co stanowi 26% wszystkich publikacji na ten temat zarejestrowanych w bazie LISTA. Do ciekawych obserwacji można też zaliczyć to, że na czasopisma tradycyjnie uznawane za najbardziej reprezentatywne dla nauki o informacji przypada blisko 56% publikacji o problematyce big data dotyczących maszynowego uczenia się oraz 45% publikacji o big data dotyczących wizualizacji informacji.

Wnioski z przedstawionego porównania trzeba jednak traktować z dużą ostrożnością zarówno ze względu na umowność kategorii „kanonicznych" czasopism nauki o informacji, jak i ze względu na relatywnie małą liczbę analizowanych publikacji i duże rozproszenie ich tematyki.

Rys. 17. Struktura tematyczna artykułów naukowych dotyczących zagadnienia big data opublikowanych w „kanonicznych" czasopismach nauki o informacji

4. Podsumowanie

Badanie piśmiennictwa zarejestrowanego w bazie LISTA potwierdziło ogólną tezę, że rosnące w nauce zainteresowanie badaniami dotyczącymi analizy wielkich zbiorów danych i jej wykorzystania w rozmaitych zastosowaniach staje się widoczne również w nauce o informacji. Liczba 381 artykułów naukowych poświęconych tej problematyce, które zostały opublikowane w czasopismach naukowych indeksowanych w LISTA, wskazuje jednak, że intensywność tego zainteresowania nie jest jeszcze bardzo duża. Niemniej jednak analiza tematyki tych artykułów potwierdziła też, że problematyka big data łączy się z kluczowymi obszarami badań nauki o informacji.

W nauce o informacji badania dotyczące big data pojawiły się w 2011 r., a więc w tym samym czasie, gdy w całej nauce nastąpił wyraźny wzrost zainteresowania tą problematyką. Podobnie jak w innych dyscyplinach, również w nauce o informacji w ostatnich ośmiu latach następował systematyczny wzrost publikacji poświęconych tej problematyce.

Artykuły, w których podejmowano tematykę big data, ukazały się na łamach aż 113 czasopism należących do szeroko pojmowanych dyscyplin informacyjnych. Równocześnie jednak wyraźnie wyodrębnia się stosunkowo nieduża grupa kilkunastu czasopism, w których w ciągu ostatnich ośmiu lat do tej tematyki powracano wielokrotnie. Są to czasopisma z zakresu: naukometrii i bibliometrii, systemów informacyjnych, informatyki medycznej, zarządzania informacją oraz zaawansowanych technologii bibliotecznych. Dostrzec też można, że o badaniach big data w problemowym obszarze nauki o informacji najczęściej publikują czasopisma specjalizujące się w pewnych jej węższych subdyscyplinach, przede wszystkim ilościowych badaniach informacji (bibliometrii, naukometrii, altmetrii), informatyce medycznej, problematyce systemów informacyjnych i wyszukiwania informacji oraz w zarządzaniu informacją. W czasopismach o szerokim profilu tematycznym publikacje na temat big data ukazują się dość rzadko.

Około 22% artykułów o big data zarejestrowanych w LISTA ukazało się w czasopismach, które w wielu badaniach uznawano dotąd za najważniejsze czasopisma nauki o informacji, najbardziej reprezentatywne dla jej pola badawczego i o wysokim wpływie na rozwój jej badań. W tej grupie najwięcej artykułów o big data opublikowano w: *Scientometrics*, *International Journal of Information Management*, *Library Hi Tech*, *Information Processing & Management* i *Information Research*. Zaskakujący jest fakt, iż w takich czasopismach jak *Journal of the Association for Information Science and Technology* czy *Journal of Documemtation,* uznawanych za główne czasopisma nauki o informacji, problematyka big data występowała dotąd sporadycznie.

Zagadnieniami big data w nauce o informacji zajmują się badacze z różnych krajów świata, są wśród nich również badacze polscy. Rozkład geograficzny afiliacji autorów artykułów o big data zarejestrowanych w bazie LISTA na ogół potwierdza z jednej strony, znaną od dawna największą aktywność ośrodków amerykańskich i brytyjskich w badaniach z zakresu nauki o informacji, w tym także związanych z tą nową technologią, a z drugiej – obserwowaną także w innych dziedzinach, gwałtownie rosnącą liczbę badań dotyczących big data prowadzonych w krajach azjatyckich, w szczególności w Chinach i w Indiach.

Piśmiennictwo dotyczące badań big data, które zarejestrowano z bazie LISTA, charakteryzuje duża różnorodność podejmowanej tematyki szczegółowej. Dominuje tematyka należąca do obszaru nauk komputerowych oraz mediów społecznych, ale do zagadnień

często omawianych należą też metadane, zarządzanie i dzielenie się wiedzą, biblioteki cyfrowe, bibliometria, w tym analiza cytowań oraz kwestie związane z informatyką medyczną i ochroną zdrowia. Z kolei w głównych czasopismach nauki o informacji na plan pierwszy wysuwają się badania bibliometryczne oraz dotyczące bibliotek cyfrowych.

Nie ulega wątpliwości, że badania dotyczące technologii big data, możliwości ich wykorzystania w usługach informacyjnych oraz rozmaitych ich aspektów społecznych będą przyciągać coraz większą uwagę badaczy nauki o informacji. Wielkość i złożoność zasobów informacji i wiedzy zgromadzonych w środowisku cyfrowym oraz stale rosnące tempo ich przyrostu wymuszają sięganie po tę nową technologię w działalności, która służyć ma skutecznemu transferowi informacji i wiedzy w społeczeństwie.

Bibliografia

Anderson, Ch. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* [online]. Wired, 6.23.08 [18.11.2018], https://www.wired.com/2008/06/pb-theory/

Chang, Y.-W., Huang, M.-H. (2012). A Study of the Evolution of Interdisciplinarity in Library and Information Science: Using Three Bibliometric Methods. *Journal of the American Society for Information Science and Technology,* 63(1), 22–33.

De Mauro, A., Greco, M., Grimaldi, M. (2016). A Formal Definition of Big Data Based on Its Essential Features. *Library Review*, 65(3), 122–135.

Ferstein, S. (2012). *Ignorance: How It Drives Science*. New York: Oxford University Press.

Friedman, A. (2018). Measuring the Promise of Big Data Syllabi. *Technology, Pedagogy and Education* vol. 27, nr 2, 135–148.

Hey, T., Tansley, S., Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research.

Intel (2012). *Peer Research Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data* [online]. Intel IT Center [18.11.2018], https://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf

Jacobfeuerborn, B. (2013). Is Big Data a Paradigm Challenge to Information Science? *Zagadnienia Informacji Naukowej*, 51 (2), 52–63.

Klous, S., Wielaard, N. (2016). *We Are Big Data. The Future of the Information Society.* Amsterdam: Atlantis Press.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*. [online]. Gartner, file No.949. 6 Feb. 2001 [18.11.2018], https://www.researchgate.net/publication/311642627_Big_Data_The_V%27s_of_the_Game_Changer_Paradigm

Mayer-Schönberger, V., Cukier, K. (2014). *Big data: rewolucja, która zmieni nasze myślenie, pracę i życie.* Warszawa: MT Biznes.

Patgiri, R., Ahmed, A. (2016). *Big Data: The V's of the Game Changer Paradigm* [online]. IEEE 18th International Conference on High Performance Computing and Communications 2016 [18.11.2018]. https://www.researchgate.net/publication/311642627_Big_Data_The_V%27s_of_the_Game_Changer_Paradigm

Reitz, J. M. (2014) *Online Dictionary for Library and Information Science* [online]. ABC-Clio [5.12.2018], https://www.abc-clio.com/ODLIS/odlis_i.aspx

Sapa, R. (2018). Reinterpretacja koncepcji użytkownika usług informacyjnych. W: *Nauka o informacji w okresie zmian: innowacyjne usługi informacyjne*. Warszawa: Wydaw. SBP, 17–26.

Saracevic, T. (2010). Information Science. In: *Encyclopedia of Library and Information Sciences. Third Edition.* Boca Raton, FL: CRC Press, vol. 4, 2570–2584, DOI: 10.108/E-ELIS3–120043704

Scopus (2018). Scopus: An Eye on Global Research [online]. Elsevier [5.12.2018], https://www.elsevier.com/__data/assets/pdf_file/0008/208772/ACAD_R_SC_FS.pdf
Sosińska-Kalata, B. (2013). Obszary badań współczesnej informatologii (nauki o informacji). *Zagadnienia Informacji Naukowej*, 51(2), 9–41.
Sosińska-Kalata, B. (2017). Kierunki rozwoju współczesnej informatologii. *Forum Bibliotek Medycznych*, 10(2), 25–46.
Ward, J.S., Barker, A. (2013). Undefined by Data: A Survey of Big Data Definitions [online]. Cornell University Library [18.11.2018], https://arxiv.org/pdf/1309.5821.pdf
White, H.D., McCaine, K.W. (1998). Visualizing a Discipline: An Author Cocitation Analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327–355.
Zhao, D., Strotmann, A. (2008). Information Science during the First Decade of the Web: An Enriched Author Cocitation Analysis. *Journal of the American Society for Information Science and Technology*, 59(6), 916–937.

Big Data (Massive Data) in Information Science

Abstract

Purpose/Thesis: The aim of the paper is to discuss main features of the phenomenon known as big data, its importance for the research issues of information science and an attempt to pre-assess the researchers' level of interest in the topic in question.
Approach/Methods: The critical analysis of literature has been used to discuss the essence of the big data phenomenon and related changes in the research model, increasingly applicable in various fields of modern science. The growing interest in big data in science is illustrated with the results of a bibliometric analysis of the literature indexed in the interdisciplinary Scopus database. The assessment of the level of interest in big data within the field of information science is based on a bibliometric analysis of the literature indexed in the domain-based EBSCO database – Library and Information Science and Technology Abstracts (LISTA).
Results and conclusions: Big data technologies can be treated as a next phase of development in computer technology and its applications in various fields of science and practice. In the environment of large data resources stored in a digital format, big data technologies provide an insight into knowledge that could not be extracted with traditional methods of information retrieval. In this sense, the afore-mentioned technologies support knowledge transfer processes occuring among people and those processes are the main focus of information science. The analysis of the literature indexed in the LISTA database confirmed that the development of big data technology and its applications is a significant challenge for information science and the interest in it is systematically growing, although it has not become very large so far. The analysis of topics of this literature also confirmed that the big data issues are related to the key areas of information science research. Most often big data research is presented in information science journals focused on quantitative information research (bibliometrics, scientometrics, altmetrics), medical computer science, information systems, information retrieval and information management. Journals with a broad thematic profile covering the whole field of information science research have been publishing papers rather rarely so far. The authors of the largest number of articles on big data in information science are affiliated to research centers in the United States, Great Britain and China. The literature on big data research in information science is distinguished with a large diversity of specific topics. Topics that belong to the area of computer science and social media dominate the field, but fairly often researchers also

discuss metadata, management and knowledge sharing, digital libraries, bibliometrics and issues related to medical computer science and health protection.
Research limitations: The research discussed in the paper is a preliminary recognition of interest in the big data phenomenon within the field of information science and it was built on the literature indexed in the LISTA database, which subject description includes the term "big data". Hence, the literature presenting the issues related to the study of large datasets where this index term was not used was not included in the study. In addition, the policy of indexing of the LISTA database, in particular the relatively small representation of journals published in languages other than English, may limit the representativeness of the results obtained for big data research related to information science issues on a global scale.
Originality/Value: To the best of the author's knowledge, the research presented in the paper is the first attempt to assess the level of interest in big data within the field of information science.

Keywords

Bibliometric study. Big data. Information science. Massive data. Research problems.

Prof. dr hab. BARBARA SOSIŃSKA-KALATA jest kierownikiem Katedry Informatologii na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego oraz redaktor naczelną czasopisma Zagadnienia Informacji Naukowej – Studia Informacyjne i członkiem Komitetu Naukoznawstwa PAN. Specjalizuje się w problematyce nauki o informacji, w szczególności organizacji wiedzy, a także ilościowych badań informacji oraz historii, teorii i metodologii nauki o informacji. Opublikowała ponad 250 prac, w tym 12 książek i ponad 160 artykułów naukowych. Do jej najważniejszych publikacji należą monografie Modele organizacji wiedzy w systemach wyszukiwania informacji o dokumentach *(1999) i* Klasyfikacja. Struktury organizacji wiedzy, piśmiennictwa i zasobów informacyjnych *(2002) oraz artykuły* Évolution des Systèmes d'Organisation des Connaissances et établissement de critères pour leur évaluation praxeologique *(2011),* Obszary badań współczesnej informatologii (nauki o informacji) *(2013),* Nauka o informacji wśród nauk o kulturze *(2017),* Książka (dokument) w środowisku informacyjnym *(2017),* The impact of the works by Paul Otlet and Suzanne Briet on the development of the epistemology of documentation and information science in Poland *(2018).*

Kontakt z autorką:
b.sosinska@uw.edu.pl
Katedra Informatologii
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski
ul. Nowy Świat 69
00-046 Warszawa