Nowa Biblioteka. Usługi, Technologie Informacyjne i Media 2019, nr 3 (34), s. 129–135
ISSN 2451-2575 (wersja elektroniczna)
GICID 71.0000.1500.0677

DOI: 10.31261/NB.2019.34.10

SPRAWOZDANIA

Karina Nabiałczyk
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytet Wrocławski
e-mail: karina.nabialczyk@uwr.edu.pl
https://orcid.org/0000-0002-0696-6450

Ogólnopolska Konferencja Naukowa „Big Data w humanistyce i naukach społecznych" (Wrocław, 22–23 listopada 2018 r.)

Tradycyjnie już w okresie jesiennym wrocławskie środowisko bibliotekoznawców zorganizowało konferencję naukową. W 2018 r. odbyła się w dniach 22–23 listopada pod hasłem „Big Data w humanistyce i naukach społecznych", a jej celem była krytyczna refleksja nad zagadnieniem big data, wymiana poglądów, doświadczeń oraz zaprezentowanie aktualnego stanu badań w tym zakresie. Za organizację wydarzenia odpowiedzialna była Aneta Firlej-Buzon, kierownik Zakładu Bibliografii i Informacji Naukowej Instytutu Informacji Naukowej i Bibliotekoznawstwa Uniwersytetu Wrocławskiego. Konferencję zainaugurował rektor uczelni, Jego Magnificencja Adam Jezierski. Następnie wręczono nagrodę za projekt i opracowanie graficzne materiałów konferencyjnych w konkursie „Studenci projektują". Otrzymała ją studentka kierunku publikowanie cyfrowe i sieciowe Patrycja Dąbrowska.

Sesję pierwszą otworzyła Barbara Sosińska-Kalata (Uniwersytet Warszawski) wystąpieniem *Badania nad danymi masowymi w nauce o informacji*, wprowadzając słuchaczy w problematykę dotyczącą nowego podejścia do gromadzenia i magazynowania cyfrowo zapisanych danych, a także do ich przetwarzania. Według B. Sosińskiej-Kalaty „big data polega na poszukiwaniu wzorców, szacowaniu prawdopodobieństwa, przewidywaniu, rekomendowaniu i wspieraniu decyzji". Prelegentka omówiła rolę dużych zbiorów danych w nauce, naukach społecznych i humanistyce. Na podstawie analizy danych z bazy Scopus wskazała,

kiedy nastąpił wzrost piśmiennictwa naukowego na ten temat. Zaprezentowała stan badań nad big data w nauce o informacji. W analizie wykorzystała bazę danych Library, Information Science & Technology Abstracts (dalej: LISTA). Badania ilościowe piśmiennictwa dostarczyły m.in. informacji na temat rozkładu chronologicznego, języka publikacji, autorów, czasopism, w których publikowano na wskazany temat, czy struktury tematycznej artykułów. Badania pokazały, że zdecydowana większość autorów piszących o danych masowych w nauce o informacji związana jest z ośrodkami badawczymi w Stanach Zjednoczonych (35%), Wielkiej Brytanii (12%) i Chinach (11%). Około 30% artykułów o big data zarejestrowanych w LISTA ukazało się w liczących się czasopismach z nauki o informacji, najwięcej w: „Scientometrics", „Journal of the American Medical Informatics Association", „International Journal of Information Management". W czasopismach uznawanych za najbardziej reprezentatywne dla całego obszaru badań (Library and Information Science) problematyka big data występowała dotąd sporadycznie. B. Sosińska-Kalata skonstatowała, że dominuje tematyka z zakresu nauk komputerowych oraz mediów społecznych, często prezentowane są także zagadnienia dotyczące metadanych, zarządzania wiedzą, bibliotek cyfrowych i bibliometrii.

Kolejny warszawski ośrodek naukowy – Uniwersytet Kardynała Stefana Wyszyńskiego, reprezentowała Katarzyna Materska. W wystąpieniu *Odpowiedzi nauk społecznych na big data* prelegentka przedstawiła różne stanowiska badaczy nauk społecznych (od entuzjastów, badaczy umiarkowanych, po nastawionych negatywnie) w odniesieniu do big data i ich uzasadnienie. Według autorki „nowy paradygmat polega na uzyskiwaniu zrozumienia danego zjawiska w oparciu o dane, a nie w oparciu o teorię, ale nie traci się jej z pola widzenia". Wskazała również różne aspekty danetyzacji rzeczywistości, w tym życia ludzkiego, np. przekazów słownych, przemieszczania się człowieka (geolokalizacja), analizy emocji czy parametrów organizmu (fenomen *quantified self*).

W następnym wystąpieniu, *Analiza korpusów literackich i naukowych za pomocą InfraNodus*, Andrzej Radomski (Uniwersytet Marii Curie-Skłodowskiej w Lublinie) omówił możliwości wymienionego w tytule oprogramowania, służącego do analizy i wizualizacji dużych korpusów tekstów – z zastosowaniem teorii grafów. Zaprezentował cechy tej aplikacji oraz podał przykłady badań wybranych korpusów literackich i naukowych, przeprowadzonych również przez siebie.

Kolejny przedstawiciel ośrodka lubelskiego, Zbigniew Osiński, przedstawił temat *Big data w praktyce badawczej humanistów – prob-*

lemy metodologiczne. Omówił kilka zagadnień (pytań) związanych z użyciem big data w badaniach, m.in. typy danych (ustrukturyzowane, częściowo ustrukturyzowane, nieustrukturyzowane), to, gdzie i w jaki sposób są wydobywane, porządkowane i przetwarzane. Zanalizował możliwość masowego kodowania „cyfrowych dzieł kultury" na wzór SemanticWeb. Na koniec postawił otwarte pytania, m.in. czy humanistyka „ilościowa", „obiektywna" będzie nauką humanistyczną.

Następny referat, *Nowe zasady organizacji informacji w środowisku big data*, wygłosił Marek Nahotko (Uniwersytet Jagielloński w Krakowie). Tradycyjną organizację informacji cyfrowej i jej narzędzi (algorytmy, wyszukiwarki, bazy danych) przeciwstawił w nim organizacji informacji w big data, gdzie gatunki organizacji informacji zyskują nowe zakresy znaczeniowe.

Sesję pierwszą zakończyło wystąpienie zespołu wrocławskich badaczy: Elżbiety Herden, Adama Pawłowskiego i Piotra Malaka, którzy wygłosili referat *Problem przetwarzania bibliografii metodami Big Data.* Punktem wyjścia ich rozważań było spojrzenie na bibliografię jako przedmiot badań humanistyki (tradycyjnej i cyfrowej) oraz wskazanie problemu badawczego: bibliografia jako przedmiot badań metodami *text mining*. Przedstawili przebieg badań nad korpusem rekordów bibliograficznych Biblioteki Narodowej w Warszawie, czyli opisów wydawnictw zwartych opublikowanych w Polsce w latach 1997–2017. Wskazali przy tym na różne problemy wynikające z automatycznego przetwarzania, m.in. na to, że obecnie dane zapisane w formacie MARC nie są przystosowane do analiz w wielkiej skali, zapis danych bibliograficznych charakteryzuje się wysoką redundancją, a wielkich bibliografii nie da się aktualizować pod względem struktury danych. Konkluzja jest jednak pozytywna: „mimo wszystko przetwarzanie wielkich bibliografii przeżywa swój renesans i ma wielką przyszłość".

W przerwie przed kolejną sesją odbyła się prezentacja Pracowni Humanistyki Cyfrowej znajdującej się w Instytucie Informacji Naukowej i Bibliotekoznawstwa. Licznie zgromadzonych zainteresowanych oprowadził kierownik Pracowni A. Pawłowski, przybliżając zasady funkcjonowania jednostki, plany naukowe oraz potencjał miejsca. Członkowie Pracowni omówili możliwości techniczne zgromadzonego sprzętu i pokazali próbki zastosowań wybranych narzędzi, m.in. okulografu do analizy odbioru materiałów graficznych.

Sesję drugą otworzyli Wiesława Osińska z Uniwersytetu Mikołaja Kopernika w Toruniu i P. Malak z Uniwersytetu Wrocławskiego. Badacze wygłosili referat *Przyszli bibliolodzy wobec wezwań Big data. Refleksje z doświadczeń personalnych*, w którym postawili pytanie,

czy edukacja bibliologów jest wystarczająca w erze big data. Ich zdaniem koncentracja uwagi w skali mikro, czyli skupienie się na pojedynczych rekordach (np. bibliograficznych książki), może zaburzać ocenę zbioru big data, dla którego właściwa jest perspektywa makro, kiedy o wyniku analiz decyduje zbiór rekordów, a na wynik ten nie wpływają pojedyncze rekordy. Autorzy zaproponowali uzupełnienie programu kształcenia o techniki przetwarzania danych wielkoskalowych, elementy statystyki, wizualizację danych, narzędzia i metody NLP.

Kolejne referaty zaprezentowali praktycy. Michał Kozak (Poznańskie Centrum Superkomputerowo-Sieciowe) w wystąpieniu *Platforma Europeana Cloud jako przykład Big Data w humanistyce* omówił problematykę agregowania metadanych udostępnianych online zasobów dziedzictwa kulturowego ze zbiorów instytucji z całej Europy. Opisał, jak jego macierzysta placówka wraz z Fundacją Europeana od 2013 r. rozwija platformę wykorzystującą technologię big data, która służy do bieżącej obsługi działań związanych z zasilaniem danymi portalu Europeana. Natomiast Maciej Jabłoński (Centrum NUKAT) w komunikacie *Katalog centralny NUKAT – największa ogólnopolska baza ustrukturyzowanych danych bibliograficznych i haseł wzorcowych – zasięg, zakres i perspektywy rozwoju* opisał zasady organizacji, struktury danych oraz projekty, w których współuczestniczy NUKAT. Zaakcentował, że dane katalogu centralnego są składową lub punktem wyjścia w tworzeniu zbiorów danych zwanych big data. Zachęcał do aktywnej współpracy między instytucjami naukowymi i jednostkami badawczymi a Centrum NUKAT. Następnie wystąpiła kierownik konferencji A. Firlej-Buzon z referatem *Wizjoner informacji naukowej w świecie danych – Eugene Garfield (1925–2017)*. Sesję drugą zamknęło wystąpienie *Informatologia a przetwarzanie dużych wolumenów danych. Case study WDIB UW* Grzegorza Gmiterka (Uniwersytet Warszawski). Referent przedstawił projekty realizowane na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego w zakresie big data, m.in. w Katedrze Technologii Informacyjnych Mediów, gdzie od 2011 r. prowadzone są zaawansowane prace nad dużymi zbiorami danych, realizowany jest program Narodowego Centrum Badań i Rozwoju dotyczący tej tematyki oraz powstało Centrum Rafinacji Informacji (spółka spin-off na Uniwersytecie Warszawskim, która zajmuje się analizą big data i jest pierwszą spółką reprezentującą nauki humanistyczne). Autor omówił również proces tworzenia nowego kierunku studiów magisterskich: zarządzanie Big Data w ramach „Programu zintegrowanych działań na rzecz rozwoju Uniwersytetu Warszawskiego" współfinansowanego ze środków Europejskiego Funduszu Społecznego w ramach POWER.

W drugim dniu konferencji, w sesji porannej wystąpił m.in. Adam Jachimczyk (Uniwersytet Warszawski) z referatem *Bazy danych z zakresu nauk humanistycznych i społecznych udostępniane w Internecie przez polskie jednostki naukowe. Wstępny rekonesans*. Prelegent zaprezentował stan udostępniania danych z zakresu nauk humanistycznych i społecznych przez polskie jednostki naukowe: szkoły wyższe, jednostki PAN, instytuty badawcze, biblioteki naukowe. W analizie wybranych przykładów zwrócił uwagę na typ gromadzonych zasobów (dane, metadane), obecność informacji o warunkach wykorzystania danych i ich dostępność w formacie ułatwiającym ponowne wykorzystanie. Przeprowadzony rekonesans badawczy pokazał, że – z pewnymi oporami – następuje rozwój repozytoriów z danymi z zakresu nauk humanistycznych i społecznych. Nie zawsze jednak towarzyszy im czytelna informacja o warunkach wykorzystania oraz udostępnianie danych w formacie ułatwiającym ich ponowne przetworzenie. Anna Łach (Uniwersytet Wrocławski) w referacie *„Roczniki Biblioteczne" w świetle badań korpusowych (analiza cytowań i frekwencji toponimów)* przedstawiła wstępne ustalenia dotyczące kierunków przepływu wiedzy, związanych z pojawiającymi się na łamach czasopisma cytowaniami dorobku naukowców pochodzących z różnych ośrodków. Wskazała m.in. typy cytowań, cytowania według płci autorów, państw czy wreszcie ranking najpopularniejszych autorów. Na podstawie wstępnych danych, określonych na bazie niewielkiego korpusu, autorka ustaliła, że „transfer wiedzy do Polski zachodzi na osi zachód-wschód". Sesję zakończył Grzegorz Zyzik (Uniwersytet Opolski) referatem *Rozgrywka i refleksja. Big Data w groznawstwie*. Intencją autora było przeanalizowanie możliwości pozyskiwania potencjału big data w tworzeniu gier wideo. Prelegent zaprezentował grę Landlord. Magnat Nieruchomości – jedną z pierwszych polskich gier wykorzystujących big data.

Ostatnią sesję, zamykającą konferencję, w całości wypełniły wystąpienia gospodarzy – pracowników i doktorantów Instytutu Informacji Naukowej i Bibliotekoznawstwa UWr. Kolejne badania empiryczne na korpusie danych zaprezentowały Kamila Augustyn i Małgorzata Góralska. W swoim wystąpieniu *„Czytanie emocji": doświadczenia z tworzeniem i znakowaniem korpusu danych na potrzeby badań nad emocjami czytelników książek* autorki postawiły następujące pytania badawcze: jakie emocje budzą książki? jakie emocje budzi czytanie książek? jaką rolę odgrywają emocje w indywidualnym i społecznym odbiorze książki? jaką funkcję pełni książka w życiu emocjonalnym człowieka? Kolejne etapy badań polegały na zbieraniu danych na potrzeby zbudowania korpusu, ich wstępnej obróbce oraz ręcznym znako-

waniu przy użyciu Inforexa – jednego z narzędzi webowych dostępnego dzięki CLARIN-PL. W trakcie prac dokonano niezbędnej klasyfikacji emocji i wskazano kategorie, według jakich można je znakować, np.: intensywność; pozytywny/negatywny wpływ; określająca ją część mowy (czasownik, rzeczownik, przymiotnik), długość zapisu (jedno-/dwuwyrazowe). Badania dowiodły m.in., że wypowiedzi wskazujące na określony stan emocjonalny nie zawsze pokrywają się z faktycznym stanem emocjonalnym badanej osoby. Następne wystąpienie dotyczyło jednej z dyscyplin nauk społecznych – prawa. P. Malak i Artur Ogurek wygłosili referat *Zastosowanie Big Data w badaniach języka prawniczego*, w którym przedstawili wyniki analizy skuteczności wyszukiwania informacji w Portalu Orzeczeń Sądów Powszechnych i rezultaty badań nad dywersyfikacją uzyskanych wyników, m.in. ze względu na rolę podmiotu w postępowaniu. Przedstawili charakterystykę profesjolektu prawniczego oraz metody analizy (m.in. NLP) zastosowane podczas badań w celu kategoryzacji orzeczeń ze względu na występujące w nich podmioty, ich role w procesach i wyniki samych procesów. Dorota Siwecka wystąpiła z referatem *Duże zbiory danych w projektach typu otwartych danych powiązanych (Linked Open Data) bibliotek narodowych w Europie w latach 2007–2017*, w którym omówiła wyniki analizy projektów realizowanych przez europejskie biblioteki narodowe w zakresie udostępniania otwartych powiązanych danych. Autorka poddała analizie m.in. liczbę realizowanych projektów, charakter udostępnianych danych (typ danych, dostępne formaty, wykorzystane schematy metadanych, źródła danych), a także informację, do jakich innych zbiorów danych dostępnych w sieci linkowane są dane biblioteczne. Ostatni z referatów, *Kolekcje średniowiecznej książki rękopiśmiennej w polskich bibliotekach cyfrowych – nowe narzędzie badawcze i jego rola w badaniach bibliotekoznawczych*, wygłosiła Kinga Brzozowska. Prelegentka podjęła próbę oceny wielkości i kompletności zbiorów, jakości źródeł informacji o zbiorach zdigitalizowanych, wartości towarzyszącego kolekcjom aparatu pomocniczego, sposobów nawigacji po zasobie oraz tempa przyrastania kolekcji wirtualnych. Przedstawiła wyniki analiz prowadzonych w latach 2013–2018. Celem referatu było również pokazanie, w jakim stopniu zbiory te mogą być przydatne w badaniach bibliotekoznawczych, a także czy na obecnym etapie rozwoju bibliotek cyfrowych właściwe jest traktowanie analizy ich zasobów na równi z tradycyjnymi metodami badań.

Spotkanie badaczy dyscyplin nauk społecznych i humanistycznych – bibliologów, historyków, bibliotekoznawców, prawników, językoznawców czy kulturoznawców – dało niepowtarzalną możliwość

szerszego spojrzenia, wykraczającego poza własne pola badawcze, na wciąż rosnący potencjał i zasoby big data. Warto zauważyć, że wśród prelegentów i słuchaczy znaleźli się przedstawiciele instytucji (m.in. Główny Urząd Statystyczny, Poznańskie Centrum Superkomputerowo-Sieciowe, Centrum NUKAT), które od lat pracują na dużych zbiorach danych i są coraz bardziej otwarte na ich udostępnianie. Wymiernym efektem konferencji będzie recenzowana publikacja, niewymiernym, a równie ważnym – możliwość wymiany doświadczeń między specjalistami różnych dyscyplin, którzy w swoich badaniach borykają się z problemami pozyskiwania, eksploracji, analizy oraz zarządzania zasobami big data.

Tekst wpłynął do redakcji 16 maja 2019 r.