

JACEK WOJCIECHOWSKI

Big Data w humanistyce i naukach społecznych,
redakcja naukowa Aneta Firlej-Buzon,
Warszawa: Wydawnictwo Naukowe i Edukacyjne
Stowarzyszenia Bibliotekarzy Polskich 2020 (Nauka –
Dydaktyka – Praktyka, 197), s. 199. ISBN 978-83-65741-56-1



Kiedy ponad sześćdziesiąt lat temu zacząłem wkraczać w obszary nauki, głównymi (ale nieprawda, że wyłącznymi) kanałami przejmowania wiedzy już wytworzonej były bibliografie – w przewadze: odbiblioteczne – oraz bezpośrednie transmisje pomiędzy pracownikami nauki. To musiało wystarczać i wystarczało (jakoś tam), tymczasem zakresy treści nieogarniętych rosły w sposób niewyobrażalny. Wraz z elektronizacją gromadzenia i wymiany danych dokonano się całkowite łąknięcie: w dotychczasowych formułach nie do opanowania. No i tak jest nadal. W dalszym rozmnażaniu tej nadprodukcji treści naukowych wszyscy bierzemy udział codziennie. Nikt nie zaprzestaje i nie przestanie. Tym bardziej że mogą być one – i zapewne w znacznym stopniu są – przydatne.

Statystyzacja oraz indeksacja tych nieprawdopodobnie licznych zasobów informacyjno-naukowych, wzbogacanych codziennie, wymagały więc jakiegoś pomysłu, w połączeniu z nową metodologią opracowania i z wykorzystaniem aparatury informatycznej. Mniej więcej w połowie ubiegłego stulecia okazało się, że stosowne szanse istnieją. Ale należało je skonkretyzować, wyłuskać, dopracować, no i oczywiście wprowadzić w życie. Ze względu na interdyscyplinarny, ale także inauguracyjny charakter całego przedsięwzięcia nazwy zaangażowanej, do tego ewentualnej naddyscypliny nie sformułowano. Do postępowania przyklejono zaś charakteryzujące określenie *Big Data*. To niekoniecznie źle, bo z tego

wywodzi się sygnał, że mamy do czynienia z niebywałą mnogością danych. Cała reszta wymaga dopowiedzeń i tak właśnie jest w tej chwili. Nowość stanowi wyszukiwanie tych mnogich danych w obszarach nauk humanistycznych oraz społecznych – jak się wstępnie okazuje: możliwe i ewentualnie użyteczne. Co jednak (na razie?) trzeba przyjąć trochę na wiarę, ponieważ dowody, argumenty i uzasadnienia do klarownych nie należą. Ale być może czas w tym zakresie też zrobi swoje: tak jak już zrobił.

Ogólnie bowiem ani porządkowanie logiczne bądź statystyczne treści, ani wyszukiwanie indeksacyjne, także w kręgu tych dyscyplin naukowych, żadną całkowitą niespodzianką nie są. Tylko stan zaawansowania wobec innych nauk jest nieporównywalny. Dotyczy to również bibliotekoznawstwa, czyli wiedzy o bibliotekarstwie, i informatologii jako wiedzy o informacji: zbudowanej na szczątkach bibliografii, w powiązaniu z informatyką.

Wchodzeniu w obszar *Big Data* w tym momencie towarzyszy omawiana publikacja, wydana w ramach serii Nauka – Dydaktyka – Praktyka. Nie mam pewności, czy ona w pełni do tej serii przystaje, natomiast wolałbym, żeby merytorycznie była zdecydowanie lepsza. Bowiem prawie połowę zamieszczonych w publikacji tekstów można było sobie darować, ewentualnie zapraszając autora lub autorów spoza Polski. Bo to może naprawdę tak być, że na razie trudno u nas o specjalistów rzeczywiście znających się na **tej** rzeczy. A wypowiadać się i dyskutować trzeba.

W tych okolicznościach wolałbym więc, żeby układ tej monografii wyglądał nieco inaczej: czytelniej. Mianowicie można by zacząć ją od wypowiedzi, które **rzeczywiście** próbują wyjaśnić to i owo w kontekście środowiska mnogich danych. I w takiej też kolejności – niech będzie, że *mojej* – odniosę się do umieszczonych w książce tekstów.

Nad redakcją całości oraz, jak rozumiem, nad strukturą tej monografii – ale tego nie napisano – czuwała **Aneta Firlej-Buzon** z Uniwersytetu Wrocławskiego, logiczne więc, że jest ona też autorką wprowadzenia, które jednak, oprócz zasygnalizowania samej nazwy, niczego nie objaśnia. Jest puste. Można było wysilić się choćby na zindeksowanie problematyki lub/oraz na informację, kto się tym zajmuje w Polsce. Ale niczego takiego nie ma. Ja zainaugurowałbym tę monografię wypowiedziami dwóch autorów z lubelskiego Uniwersytetu Marii Curie-Skłodowskiej, ale w innej kolejności, niż to jest: ze względu na dążenie do wyjaśnień. Oto **Andrzej Radomski** sygnalizuje, że z powodu obecnej eksplozji danych oraz zalewu tekstów cyfrowych trzeba było odstąpić od dominującej uprzednio roli interpretacji danych na rzecz wykorzystywania nowych narzędzi oraz programów. No i z tego narodziła się nowa dziedzina

(megadzielnina?) nauki, nazywana właśnie *Data Science* albo *Machinelearning*. Zajmując się (także) psychologią poznania, żywię tylko nadzieję, że nie jest to wjazd na autostradę donikąd. Bo jednak nauki o ludzkim myśleniu przestrzegają przed nadmiernym zaufaniem w stosunku do maszyn. A zwłaszcza wobec *sztucznej inteligencji* (*artificial intelligence*). Takiego sformułowania w wypowiedziach o ambicjach naukowych lepiej nie używać. Niech już będzie – lepiej – *humanistyka cyfrowa*, skupiająca nad mnogimi danymi reprezentantów dyscyplin dotychczasowych, a w opracowaniu wykorzystująca metody statystyczne. Od luźnych zestawień (program Tagul) poprzez statystyczną strukturyzację (Vayant) po zwizualizowaną sieć interaktywną (Infra Nodus). Czy i jak to zagra, to dopiero zobaczymy. Ale prawdą jest, że czekać nie ma na co. Z kolei nie ma też gwarancji, że gdzieś nie czyha wywrotka. Zwłaszcza jeżeli szanse kontroli rzeczywiście okażą się złudne.

W podobnym tonie wypowiada się **Zbigniew Osiński**, też z Uniwersytetu Marii Curie-Skłodowskiej, dopisując różne komentarze uzupełniające oraz wątpliwości. Ale główny kłopot polega na tym, że ten tekst jest kiepsko napisany. Pełno w nim rozmaitych ułomności referencyjnych: *autor przeprowadził, zaznaczono...* Główna autorska refleksja snuje się około (częściowej?) niewydolności humanistyki naukowej, na co lekarstwem może być humanistyka cyfrowa. Ale czy jest albo czy będzie? W sumie potrzebne jest wyłuskiwanie informacji w trybie różnych analiz mediów, z wykorzystaniem programów – które niby są, lecz (jeszcze?) ich (jako struktur) nie ma – interpretujących dane, a nie same teksty. Trzeba powiedzieć, że nie brzmi to dobrze. Tym bardziej podzielam autorskie niepokoje dotyczące szukania sposobów analizowania opinii oraz emocji. Jak też określania relacji pomiędzy zapisem i sensem oraz inne wątpliwości, a jest ich niemało. Stąd tak znaczne i pilne zapotrzebowanie na programy praktyczne. Osiński pisze też o tym, co inni pomijają, mianowicie, że komputery nie radzą sobie z alegoriami oraz w ogóle z pośredniością znaczeń. To dlatego cały obszar pośredniczenia w komunikacji został przez rozmaitych *teoretyków* odcięty od literatury i sztuki. Dla informatologów powieści, bajki czy poezje nie istnieją. W co brną bez opamiętania również pseudobibliotekoznawcy, także z grona współautorów tej monografii. Dlatego zidentyfikowanie przez Autora konkretnie tego problemu uważam za ważne.

I – moim zdaniem – tyle dobrego. Pozostałe wypowiedzi operują ogólnikami, nie serwują objaśnień użytecznych ani produktywnych komentarzy, więc chyba na taki tom było u nas za wcześniej. Albo dobrane grono autorskie nie w pełni sprostało zadaniu.

Krystyna Materska z Uniwersytetu Śląskiego przekonuje, że cyfryzacja, a nawet danetyzacja (od: danych) materiałów naukowych to już rzeczywistość. Można bowiem gromadzić ogrom danych bez porządkowania, co zresztą oznaczałoby jednocześnie zagładę ekspertów. Pisać można różne głupstwa... Opinia jest wzięta chyba z sufitu, skoro zaraz potem pojawia się konkluzja, że dane trzeba jednak przetwarzać. Trudno wobec tego nie zapytać: kto?, co? oraz jak? I domyślam się, że odpowiedź brzmi – sieć, maszyny, odpowiednio wypracowane programy. Które już są lub niebawem będą. Chociaż nie wiadomo, jak dalece skuteczne przy pomieszaniu metod ilościowych oraz jakościowych. A takich koncepcyjnych ogólników jest w tym tekście moc. Kiedy pada autorska sugestia, że ma miejsce rafinacja informacji, niedostrzegalnych wprost – tworzą się tzw. *stupy* – to ewentualnie należałoby uwierzyć, że to dobrze, ale kto nad tym panuje i czerpie z tego pożytek? A także: co dalej? Objasnienia w tym tekście są mętne, poza prostoduszną konkluzją, że należy wypracować koncepcje informatyczne dla socjologów. Humanistów Autorka już nie wymienia, mimo że jej instytut jest z pochodzenia bibliotekoznawczy, a nie socjologiczny. Ale obok wątpliwości, słusznych bądź nie, jest w tym tekście coś, co zaniepokoiło mnie szczególnie, zwłaszcza jeżeli podobnie myślą również inni. Otóż zdaniem Autorki odtąd sens wiedzy naukowej sprowadzi się do ustalania **CO**, zamiast jak dotychczas **DLACZEGO**. Nie tylko według mnie¹ byłoby to całkowite odejście od obszarów i metod, do których nauka jest (była?) przypisana. Dopóki więc nie zafiksują czegoś takiego odpowiedzialne i autorytatywne grona naukowe, doradzałbym w tym zakresie fundamentalną ostrożność.

Na klarowniejszy nie wygląda również tekst **Marka Nahotki**, napisany również w celu zaproponowania objaśnień, ale są one ogólnikowe i mętne. Na to nakładają się jeszcze stylistyczne dziwaczności, w rodzaju *przedstawiona zostanie, omówiona będzie*. To dodatkowo utrudnia zrozumienie. Jest tam też garść opinii wątpliwych lub niezgodnych z prawdą, jak chociażby ta, że każde ludzkie działanie wiąże się z jakąś formą komunikacji. Bieganie, spanie też? Nie można wypisywać takich rzeczy. Ani że komunikacja zawsze jest mediowana (?). Bądź że materialność algorytmiczna to też kodowanie ludzkich intencji. **Intencji** na pewno nie. Są poza tym sformułowania czysto hasłowe, np. że *Big Data* wpływa na organizację informacji. A coś bliżej? Oraz że w przypadku dużych zbiorów danych jest/będzie możliwe zautomatyzowane wyszukiwanie informacji. Tu także przydałaby się garść detali, tym bardziej że w końcu o tym właśnie jest cała monografia. Autor twierdzi, że organizacja informacji opiera

¹ H. Grabowski, *Czym jest nauka?*, „Forum Akademickie” 2020, nr 9, s. 71.

się na przesłankach wrodzonych – komu?, co to znaczy? – na co jednak żadnego dowodu nie dostarcza. Jego zdaniem współczesne formy wyszukiwania informacji mają charakter typowy i powtarzalny. Zapewne ma rację, ale to nic nowego: prawdopodobnie zawsze miały. Jednocześnie dopowiada (ale to chyba sprzeczność), że środowisko *Big Data* wpływa na organizację informacji. Już nie zapytam – jak. Do tego dolepią jest ogólnikowa opinia, mówiąca o tym, że biblioteki straciły oto monopol na ofertę informacyjną, ale to nieprawda, bo go nigdy nie miały. Przewagę owszem, ale monopolu – nie. Pojawia się jeszcze sugestia, że teraz na biblioteki czekają nowe wyzwania. Jednak żadne konkrety, nazwy ani wypowiedzi nie padają. Wszystko to w proweniencyjnym kontekście instytutu, który jeszcze niedawno miał w nazwie termin *bibliotekoznawstwo*.

Jeżeli usunie się na bok wszelkie wątpliwości, to w tym tekście niewiele zostanie. Autor ma zapewne rację, sugerując, że gatunki informacji mogą być rozpoznawalne na podstawie schematów mentalnych oraz że wyszukiwarki nie są neutralne za sprawą algorytmów. Mimo to twierdzi, że zautomatyzowane wyszukiwanie informacji w dużych zbiorach jest i będzie możliwe. Jakkolwiek dopowiedzenie, że bez hipotez, budzi mój niepokój. Bo to był dotychczas fundament nauki, ale ja należę do pokolenia minionego. Natomiast ogólnie rzecz biorąc, nadmiaru świeżych opinii ta wypowiedź nie wprowadza: jest pustawa.

Dalsza część monografii, oprócz prezentacji sylwetek dwojga prekursorów mnogodanowego postępowania, zawiera próby praktycznego sprawdzenia formuły *Big Data* oraz (przynajmniej) przykładowego wykorzystania. Ale zabieg nie jest udany. Tym bardziej że ubarwiony pustymi zwrotami śródtytułów: *praktyka, konteksty*. Jaka praktyka? Jakie konteksty? To nie znaczy nic.

Piotr Malak oraz **Artur Ogurek** z Uniwersytetu Wrocławskiego wzięli pod lupę portal orzeczeń sądów powszechnych, w sumie 295 tys. orzeczeń, ale w rzeczywistości to raptem 1060 dokumentów – z występującym hasłem *biblioteka*. Domyślam się, dlaczego nie z hasłem *pies* albo *piłka nożna*, natomiast trudno zrozumieć, w jakim celu. Poza tym akurat tyle to ja sam potrafiłbym rozeznąć w pojedynkę, pomiędzy kolejnymi wizytami w piwiarni, więc gdzie tu i po co *Big Data*?

Autorzy podjęli przeszukiwania, za kryteria wyszukiwawcze uznając: tematykę, wyróżnione frazy oraz odwołania. Ostateczne rezultaty są takie, że *biblioteka* pojawia się w tych dokumentach jako *pozwana* w 14% przypadków, jako *powód(ka?)* w 3%, a jako *uczestniczka* postępowań – jednak nie wiadomo jakich – w 83% przypadków. Muszę jednak zapytać: co miałyby z tego wynikać? Że biblioteki nie atakują swoich użytkowników? Jaki jest pożytek i w dodatku dlaczego inny niż do wyłuskania w trybie

tradycyjnym. Może nastawiam się zbyt opornie, ale korzyści nie widzę, a już na pewno nie w sugestiach (które tam są) dotyczących zmiany całego postępowania naukowego. Bez przesady!

Z kolei **Adam Pawłowski** razem z **Elżbietą Herden**, przy wsparciu **Piotra Malaka** – wszyscy też z Uniwersytetu Wrocławskiego – zaproponowali porównanie możliwości przetworzeń bibliograficznych w praktyce minionej z tym, co dałoby się uzyskać z zasobów zgromadzonych obecnie. Za przykład posłużyły lata 1997–2017 i półmilionowy zasób publikacji. Pod uwagę zostały wzięte jednak wyłącznie tytuły i, szczerze mówiąc, nie do końca rozumiem, po co. Bo nawet uczeń szkoły podstawowej lepiej lub gorzej rozumie, że zupełnie inaczej konstruuje się wypowiedzi werbalne, a inaczej tytuły publikacji. W szczególności wydawnictw zwartych. W każdym razie autorzy donoszą, że w tytułaturze polskich wydawnictw zwartych ze wskazanych dziesięciu lat zdecydowanie przeważają rzeczowniki, podczas gdy w bezpośredniej komunikacji interpersonalnej bezdyskusyjna jest dominacja czasowników. Przyznaję, że ta wiadomość nie powaliła mnie na łopatki. Jak również nie wiem, komu i do czego jest potrzebna. Z autorskich rozważań wynika jeszcze, że optymalny tytuł książki powinien składać się z dwóch–trzech wyrazów. Więc ani z jednego, ani z wielosłowa. Nie ma powodu zaprzeczać, ale już od dawna utrzymuje się taka mniej więcej świadomość, toteż trudno mówić o jakiegokolwiek odkrywczości.

Absolutnie kiepskie jest natomiast opracowanie **Anny Łach** – również z Uniwersytetu Wrocławskiego – wybiórczo analizujące rejestry przywołań w „**Rocznikach Bibliotecznych**”, jednak tylko z lat 1957 i 2017. W sumie to raptem 24 teksty, a więc *Data* są nie bardzo *Big*. Autorka poza tym dopowiada, że to było i jest czasopismo rozpoznawalne wśród bibliologów, ale to określenie jest nieprawdziwe, bo przez wiele lat takiego nie było, a i dzisiaj też mam wrażenie, że już nie ma. Na dodatek stylistycznie wypowiedź jest katastrofalna: *objęto, poczyniono, przygotowano, przeanalizowano...*

Zestawienia cytowań z obu sygnalizowanych lat trudno wzajemnie porównywać, nie wiadomo bowiem, jak dalece są reprezentatywne. W każdym razie odniesienia do płci przywoływanych autorów/autorek uważam za nonsens, tym bardziej że w niemałym stopniu ustalenie tej płci okazało się niemożliwe. Z innych porównań – zakładam, że wiarygodnych – wynika, jakoby w 2017 roku dominowały przywołania rodzime aż w 85%, wobec mniej licznych – 65% – w 1957 roku. No i co z tego? Autorka jest zdania, że obecnie dominują zakotwiczenia głównie lokalne oraz że przywołaniom towarzyszą długie, często kilkudziesięcioletnie zwłoki. No dobrze, ale powtórzę: co miałyby z tego wynikać? Zwłaszcza

przy ewidentnie historycznych ciągach „Roczników”. Bardzo przepraszam, nie potrafię wskazać żadnego powodu, dla którego ten tekst został opublikowany.

A już całkowitym nieporozumieniem jest wypowiedź **Małgorzaty Baki** z Uniwersytetu Mikołaja Kopernika w Toruniu, odnosząca się do zasobów muzealnych. Nieczytelna i zbudowana z samych ogólników.

Poza tym są tam jeszcze dwa teksty, charakteryzujące, na różnych zresztą poziomach ważności, dokonania dwóch osób uwikłanych w koncepcje oraz w praktyki statystycznego opracowania oraz wykorzystywania informacji. Społecznych, humanistycznych i jakichkolwiek. Otóż **Bożena Koredczuk** z Uniwersytetu Wrocławskiego przypomina **Marię Czarnowską** (zmarła w 2001 roku) z Biblioteki Narodowej i jej ogromny wkład w rozwój polskich rejestrów edytorskich oraz częściowo bibliografii. Rocznik „Ruch Wydawniczy w Liczbach”, ważny dla kilku zawodowych specjalności, w tym też bibliotekarskiej, wydawany przez Bibliotekę Narodową przy współpracy GUS, był przez wiele lat głównym rezultatem jej działalności. Początkowo wypracowywany ręcznie – nigdy nie pytałem o *kuchnię* – ale z czasem w końcu zdigitalizowany. A materiał do opracowania był bezdyskusyjnie obfity. Czarnowska (jak zapamiętałem) bardzo przy tym narzekała, że musi obchodzić się bez wydawnictw drugiego obiegu.

Wypowiedź ta jest interesująca oraz, według mojej oceny, trafna. Z jednym wszakże zastrzeżeniem. Mianowicie Czarnowska nie była w Bibliotece Narodowej jedyną osobą, która posługiwała się narzędziami statystycznymi w stosunku do danych mnogich. Równolegle robił to też **Jerzy Maj**, a ponieważ w zakresie bibliotekarstwa² – przynajmniej jedno zdanie Jurkowi się należało. W końcu w książce firmowanej przez SBP i udającej, że jakiś związek z bibliotekarstwem ma.

Eugene Garfield, indeksator, wizjoner i postać z zupełnie innej naukowej półki, był oczywiście do scharakteryzowania znacznie trudniejszy. Podjęła się tego redaktorka całości, **Aneta Firlej-Buzon** z Uniwersytetu Wrocławskiego. Jednak rezultat jest połowiczny. Autorska relacja wydała mi się nadmiernie ogólnikowa, dodatkowo zamazana nienaturalnym szykiem wypowiedzi i dziwaczną stylistyką. Gdybym w swoim czasie nie czytał wypowiedzi Garfielda w oryginale, w tej prezentacji odkryłbym mniej, niż należy. Za dużo jest zwłaszcza odniesień do jego sukcesów komercyjnych, które oczywiście były, ale w tym tomie akurat nie to jest najważniejsze. Z naszego punktu widzenia liczy się głównie jego

² „Biblioteki Publiczne w Liczbach” 1972–2015; *Statystyka w bibliotece i jej otoczeniu*, Warszawa 2007; *Elementarz MAK-a dla bibliotekarzy*, Warszawa 1999.

gigantyczny wkład w indeksowanie. Autorka ma oczywiście rację, przypisując mu inspirację Impact Factor, z kontynuacją w postaci Arts and Humanities Citation Index oraz Social Science Citation Index. Czy jednak – biorąc pod uwagę tytuł tomu – nie dało się tego lekko rozwinąć oraz wzbogacić w jakieś szczegóły? Pytam, bo stopień trudności zadania, zwłaszcza na początku, chyba nie został dostatecznie uświadomiony. Przecież pierwsze podejmowane przez niego próby indeksowania dużych zbiorów danych były realizowane z użyciem kart perforowanych... Proszę spróbować! Dopiero potem pojawiły się komputery.

Trochę więc cała ta opinia wypada zdawkowo. Z kolei akurat dla mnie okolicznością ważną jest fakt, że z Garfieldem przez dłuższy czas współpracował taki tuz bibliotekoznawstwa jak **Jesse Shera**³. Późdaniowa o tym wzmianka to nieco za mało.

Tak więc Wydawnictwo SBP zaoferowało knot, który kupiłem, z ciekawym tytułem oraz ufając firmie edytorskiej i jakości serii. To był jednak błąd. Bo sama seria utwierdzała swoją wartość dawno, kiedy bibliotekoznawstwo istniało u nas jako silna dyscyplina naukowa, uprawiana i wykładana samodzielnie w kilkunastu ośrodkach akademickich. Także w UAM. I to od tamtejszego instytutu rozpoczęło się spuszczenie po brzytwie centrów oraz samej dyscypliny. Na początku przyklejono bibliotekoznawstwo do ułamkowego obszaru nauki o książce, z fikcyjną nazwą *bibliologia*. Ale ta nauka zamarła, po niewątpliwym zresztą osiągnięciu⁴, i odtąd zaczęło się wypieranie bibliotekoznawstwa z ośrodków akademickich przez naukę o informacji, z niezgodnym z prawdą stwierdzeniem, że bibliotekarstwo oraz informacja to jedno i to samo. W rezultacie dla tysięcy bibliotek prawie nikt u nas kadr nie kształci na poziomie wyższym i nie ma też naukowej refleksji.

Ale być może dokona się odnowienie, bo w końcu potrzeba jest matką wynalazków. Tak jest przecież w UAM, na Wydziale Pedagogiczno-Artystycznym w Kaliszu. Wypada życzyć powodzenia, jak i utrzymania czy rozruchu akademickiego kształcenia bibliotekarzy również gdzie indziej. W Polsce istnieje kilkadziesiąt tysięcy bibliotek – czy to słaby argument?

Tekst wpłynął do Redakcji 12 października 2020 roku.

³ J. Shera, *Sociological foundations of librarianship*, New York 1970.

⁴ *Encyklopedia książki*, t. 1–2, Wrocław 2017.