Bartłomiej Włodarczyk
bm.wlodarczyk@uw.edu.pl
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytetu Warszawskiego
http://dx.doi.org/10.18778/8142-336-6.10

OPRACOWANIE W CHMURZE CZY CHMURY NAD OPRACOWANIEM? AUTOMATYCZNE INDEKSOWANIE DOKUMENTÓW A BIBLIOTEKI

Abstract: The paper presents recent research in the field of automatic indexing of text documents, inter alia, in libraries, and the attitudes of Polish academic librarians towards the computerization of the subject cataloging. The methods of literature review and survey were used along with the analysis of Polish academic curricula in the field of library and information science. The article demonstrates on several examples that the similarities in document layout and the topical diversity or homogeneity are the key factors in the computerization of cataloging. The survey conducted amongst Polish subject indexing specialists from academic libraries shows that they have highly limited knowledge about automatic indexing. The results are then compared with the findings of the study on German- and English-speaking librarians' opinions about automatic subject indexing. They are similar to the outcomes of the previous research by Alice Keller into the attitudes of, among others, the English-speaking subjects.

Słowa kluczowe: automatyczne indeksowanie, semi-automatyczne indeksowanie, opracowanie rzeczowe, opracowanie formalne, biblioteki akademickie

Wstęp

Pierwsze komputery pomagały w wykonywaniu rutynowych, powtarzalnych czynności takich jak na przykład złożone obliczenia. Wprowadzenie tych maszyn jako narzędzi wspomagających pracę ludzką nastąpiło na szerszą skalę w latach siedemdziesiątych XX w. wraz z pojawieniem się na rynku komputerów osobistych. Rozwój techniczny prowadził do coraz powszechniejszego ich stosowania w różnych dziedzinach. Obecnie sprzęt kontrolowany za pomocą algorytmów komputerowych może wykonywać działania, które jeszcze dekadę temu wydawały się być zarezerwowane wyłącznie dla człowieka z jego zdolnością szybkiego uczenia się i przystosowywania do różnych warunków pracy. Przykładem takiej czynności jest prowadzenie samochodu, które wymaga uwzględnienia wielu elementów i relacji między nimi oraz szybkiego

podejmowania decyzji. Prace nad samochodami autonomicznymi, prowadzone między innymi przez firmę Google, są przykładem wykorzystania odpowiednich algorytmów do rozwiązywania coraz bardziej skomplikowanych problemów[1]. Jednym z interesujących tematów badawczych, zyskującym szerokie zainteresowanie również poza światem nauki[2], jest kwestia możliwości komputeryzacji zawodów rozumiana jako możliwość całkowitego zastąpienia pracy ludzkiej przez maszyny obliczeniowe. Badania Carla Benedikta Freya oraz Michaela A. Osborne'a z Uniwersytetu w Oksfordzie miały na celu zbadanie tego zagadnienia w aspekcie podatności poszczególnych zawodów na komputeryzację. Według angielskich naukowców proces ten może dotknąć różnych profesji, w których nie występują bariery niemożliwe do przezwyciężenia z punktu widzenia informatyki. Zidentyfikowali oni trzy elementy, które na obecnym etapie rozwoju tej dziedziny, stanowią granicę komputeryzacji zawodów. Są to[3]:

- wysoki poziom percepcji i manipulacji;
- inteligencja kreatywna;
- inteligencja społeczna.

Stopień obecności tych elementów w zadaniach wykonywanych przez ludzi określa granicę komputeryzacji różnych profesji. Badacze pokazali jedynie potencjalny zakres zmian odnośnie do 702 zawodów, nie określili jednak kiedy one nastąpią. W poniższej tabeli (tab. 1) zaprezentowano prawdopodobieństwo komputeryzacji wybranych zawodów obliczone przez Freya i Osborne'a. Prawdopodobieństwo równe „0" oznacza, że dany zawód nie zostanie zastąpiony przez komputery, natomiast prawdopodobieństwo równe „1" oznacza, że profesja niemal na pewno zostanie zastąpiona przez komputery.

W tabeli na pozycji numer 360 z prawdopodobieństwem wynoszącym 0,65 znajduje się zawód bibliotekarza. Jego pozycja odzwierciedla zróżnicowany zakres obowiązków pracowników biblioteki, z których część podlega łatwiejszej komputeryzacji, część zaś należy do wymienionej powyżej grupy zadań nie poddających się temu procesowi. Jednym z podstawowych zadań bibliotekarzy jest zapewnienie dostępu do dokumentów poprzez odpowiednie

[1] C.B. Frey, M.A. Osborne, *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, „Technological Forecasting and Social Change" 2017, vol. 114, s. 255.

[2] Dowodem na to może być medialne zainteresowanie w postaci licznych artykułów omawiających przedstawione poniżej badania C.B. Freya i M.A. Osborne'a. Zob.: E. Pofeldt, *Study: Will A Robot Do Your Job Some Day?*, https://www.forbes.com/sites/elainepofeldt/2014/02/26/will-r2-d2-snag-your-job/#716f4ad51897, [dostęp: 6.04.2017] oraz N. Stylianou [et al.], *Willa Robot Take Your Job?*, http://www.bbc.com/news/technology-34066941, [dostęp: 6.04.2017].

[3] C.B. Frey, M.A. Osborne, *op. cit.*, s. 261.

opracowanie formalne i rzeczowe zbiorów. Ten rodzaj pracy wydaje się być podatny na komputeryzację, w odróżnieniu na przykład od działań polegających na aktywizacji kulturalnej środowiska lokalnego.

Tabela 1

Prawdopodobieństwo całkowitej komputeryzacji wybranych zawodów obliczone przez C. B. Freya i M. A. Osborne'a.

Numer pozycji z oryginalnej tabeli	Prawdopodobieństwo	Zawód
1.	0,0028	Terapeuci rekreacyjni
13.	0,004	Choreografowie
360.	0,65	Bibliotekarze
415.	0,76	Archiwiści
616.	0,96	Pomocnicy biblioteczni biurowi
702.	0,99	Telemarketerzy

Źródło: C.B. Frey, M.A. Osborne, *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, „Technological Forecasting and Social Change" 2017, vol. 114, s. 269–278.

W artykule sformułowano pytanie o obecny rozwój automatycznego indeksowania oraz stosunek bibliotekarzy do możliwości automatyzacji opracowania. Posłużono się metodą analizy i krytyki piśmiennictwa, przejrzano programy i plany zajęć kierunków kształcenia związanych z bibliologią i informatologią na polskich uczelniach oraz przeprowadzono wstępne badanie ankietowe dotyczące opinii bibliotekarzy na temat komputeryzacji indeksowania tekstowych zbiorów bibliotecznych. Dalsza część artykułu została podzielona na trzy części. W pierwszej przedstawiono przyczyny zainteresowania automatycznym opracowaniem w bibliotekach, rodzaje oraz przykłady automatycznego indeksowania zasobów, w drugiej zaprezentowano wyniki badania opinii polskich bibliotekarzy z bibliotek akademickich oraz porównano je z wynikami badania przeprowadzonego wśród bibliotekarzy anglo- i niemieckojęzycznych. W zakończeniu przedstawiono podsumowanie problematyki przedstawionej w artykule oraz zarysowano możliwości dalszych badań.

Przyczyny zainteresowania automatycznym indeksowaniem, jego rodzaje oraz przykłady zastosowania

Istnieje kilka przyczyn zainteresowania automatycznym indeksowaniem zbiorów bibliotecznych. Wynikają one w dużym stopniu z ograniczeń i wad opracowania manualnego, wobec którego podstawowe zarzuty podsumował Pierre de Keyser stwierdzając, że indeksowanie manualne[4]:

- jest wolne;
- jest drogie;
- jest niewystarczająco szczegółowe;
- niekoniecznie prowadzi do lepszego wyszukiwania;
- jest oparte na słownictwie kontrolowanym, które jest przestarzałe;
- jest oparte na słownictwie kontrolowanym, które jest skupione na dokumencie;
- nie prowadzi do spójnego opracowania.

Wydaje się, że część wymienionych problemów wiąże się silniej z niedostatecznym stopniem wyszkolenia pracowników, niż z trybem pracy. Warto zwrócić jednak uwagę na dwie kwestie związane z manualnym indeksowaniem: koszty oraz brak spójności.

Pierwszy z wymienionych zarzutów, dotyczący kosztów opracowania dokumentów, jest związany przede wszystkim ze wzrastającą liczbą zasobów, które należy zindeksować i zapotrzebowaniem na szybkie dostarczenie informacji o zbiorach. Dysponujemy stosunkowo niewielką ilością informacji dotyczących kosztów poszczególnych procesów bibliotecznych w konkretnych bibliotekach, przy czym interesujące są przede wszystkim najnowsze tego typu dane pochodzące ze współczesnych skomputeryzowanych bibliotek. Na przykład w roku finansowym 2015[5] do bazy MEDLINE wprowadzono ponad 806 tys. opisów, a średni koszt opisu jednego artykułu wynosił 9,40$[6]. Kate-Riin Kont przedstawiła dokładniejsze dane dotyczące kosztów poszczególnych etapów katalogowania w bibliotekach Estońskiej Akademii Muzyki i Teatru (EAMT) oraz Uniwersytetu Technicznego w Tallinie (UTT) w latach 2012–2013. Biblioteki te znacznie różnią się, jeśli chodzi o gromadzone zasoby.

[4] P. de Keyser, *Indexing: From Thesauri to the Semantic Web*, Oxford 2012, s. 40–47. Autor omawia każdy z wymienionych problemów związanych z manualnym indeksowaniem.

[5] Federalny rok finansowy jest liczony od 1 października do 30 września.

[6] D. Demner-Fushman, J. Mork, *NLM Medical Text Indexer Technical Report to the LHNCBC Board of Scientific Counselors April 2016*, Bethesda 2016, s. 4, https://lhncbc.nlm.nih.gov/system/files/pub9359.pdf, [dostęp: 6.04.2017].

Pierwsza z nich gromadzi głównie druki muzyczne i dokumenty audiowizualne, druga zaś głównie książki[7]. Na początku swojego artykułu autorka zaznaczyła, że obecnie konieczność uzasadnienia ponoszonych kosztów stała się istotnym elementem pracy kierowników bibliotek[8]. Jest to ważne stwierdzenie, ponieważ wiąże się z potrzebą obniżania kosztów, a automatyczne indeksowanie mogłoby pomóc w osiągnięciu tego celu. W bibliotece UTT średni koszt katalogowania pozycji w języku obcym wynosił 5,21€, a dokumentu w języku estońskim – 0,95€. Ta znaczna różnica w kosztach indeksowania wynikała z tego, że publikacje w języku estońskim są opracowywane najpierw przez Estońską Bibliotekę Narodową (EBN), a badana biblioteka kopiuje te opisy do swojego katalogu. W przypadku dokumentów zagranicznych największe koszty katalogowe były związane ze sporządzeniem opisu bibliograficznego (37,24% wszystkich kosztów), a następnie z klasyfikacją i przydzieleniem sygnatury (18,23%) oraz ze sporządzeniem opisu przedmiotowego (16,31%). Nieco inaczej koszty katalogowania rozkładały się w bibliotece EAMT, która nie ma możliwości kopiowania opisów z EBN. Koszt skatalogowania dokumentu w języku obcym wynosił w tej bibliotece 5,22€, a publikacji w języku estońskim – 3,28€. Zdecydowaną większość kosztów katalogowych pochłania w bibliotece EAMT tworzenie opisów bibliograficznych (56,32% dla dokumentów zagranicznych i 35,98% dla dokumentów estońskich). Różnice między obiema bibliotekami wynikają z odmienności gromadzonych zbiorów[9]. Opisane koszty jednostkowe należy przemnożyć przez liczbę nowych dokumentów wprowadzanych do katalogu biblioteki. Koszty katalogowania są zatem istotnym elementem wydatków bibliotek i każda możliwość ich ograniczenia, bez straty jakości opracowania, powinna być brana pod uwagę. Jedną z nich może być przynajmniej częściowa automatyzacja opracowania zbiorów.

Drugi ze wspomnianych problemów wiąże się ze spójnością opracowania zbiorów. W artykule zatytułowanym „Cataloging must change!" („Katalogowanie musi się zmienić!") Dorothy Gregor i Carol Mandel postulowały, odwołując się do wyników badań Marcii Bates oraz Lois Mai Chan, żeby bibliotekarze nie skupiali się zbytnio na zachowaniu spójności opracowania rzeczowego, ponieważ jej osiągnięcie nie jest możliwe. W artykule wspominano o spójności na poziomie 10–20%[10]. Odpowiedzią na to stwierdzenie był artykuł autorstwa Thomasa Manna, w którym skrytykował on przedstawiony powyżej

[7] K.-R. Kont, *How Much Does It Cost to Catalog a Document? A Case Study in Estonian University Libraries*, „Cataloging & Classification Quarterly" 2015, vol. 53, issue 7, s. 836.

[8] *Ibidem*, s. 826.

[9] *Ibidem*, s. 845–847.

[10] D. Gregor, C. Mandel, *Cataloging Must Change!*, „Library Journal" 1991, vol. 116, issue 6, s. 46.

postulat, podkreślając, że wynika on z niezrozumienia wcześniejszej literatury przedmiotu, która jego zdaniem pokazywała brak spójności w przypadku nie występowania żadnej formy kontroli słownictwa. Autor podkreślał również, że akceptacja braku spójności w opracowaniu prowadzi do podważenia głównego celu katalogowania, którym jest gromadzenie dokumentów na ten sam temat pod identycznymi punktami dostępu. Jego zdaniem brak zrozumienia tej podstawowej zasady powoduje, że nie ma ograniczeń, jeżeli chodzi o spadek jakości indeksowania. Jej podważenie umożliwia jednak zwiększenie liczby opracowywanych zasobów[11]. Analiza wyników badań prowadzi do wniosku, że spójność opracowania jest bardzo zróżnicowana. Dla oceny jakości indeksowania i jego przydatności spójność nie jest jednak jedyną zmienną, którą należy brać pod uwagę. Spójne może być zarówno poprawne, jak i błędne indeksowanie, co nie oznacza, że jest ono równie dobre. Jednocześnie rezultatem poprawnego indeksowania jest wysoka spójność[12]. Podsumowując, szczególnie istotny, jeśli chodzi o chęć wdrożenia indeksowania automatycznego, jest koszt opracowania zbiorów. Brak spójności manualnego indeksowania również może być argumentem za jego komputeryzacją, trzeba jednak pamiętać o złożoności oceny tego zjawiska.

Istnieje kilka sposobów automatycznego tworzenia metadanych, które wymagają nieco odmiennego podejścia i charakteryzują się różnym stopniem komplikacji. Cztery podstawowe to[13]:

- wyodrębnianie metatagów;
- ekstrakcja treści;
- automatyczne indeksowanie;
- zewnętrzna autogeneracja metadanych.

W pierwszym przypadku pola metadanych są uzupełniane przez program komputerowy metadanymi zawartymi w danym dokumencie lub z nim bezpośrednio powiązanymi[14]. Ekstrakcja treści polega na wyodrębnianiu metadanych

[11] T. Mann, *„Cataloging Must Change!" and Indexer Consistency Studies: Misreading the Evidence at Our Peril*, „Cataloging & Classification Quarterly" 1997, vol. 23, issue 3–4, s. 40–42.

[12] K. Golub i in., *A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval*, „Journal of the Association for Information Science and Technology" 2016, vol. 67, issue 1, s. 6.

[13] J. Park, A. Brenza, *Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art*, „Information Technology & Libraries" 2015, vol. 34, issue 3, w różnych miejscach. Jung-ran Park i Andrew Brenza wymieniają jeszcze eksplorację tekstu i danych oraz społecznościowe tagowanie. Wyłączono je z zaprezentowanego wyliczenia, ponieważ nie są one rodzajami automatycznego generowania metadanych, lecz odpowiednio przykładem dziedziny, której osiągnięcia są wykorzystywane w tym procesie oraz działalności ludzkiej na dużą skalę.

z treści opracowywanego zasobu informacyjnego[15], natomiast automatyczne indeksowanie zakłada przypisanie tak wyodrębnionych danych do kontrolowanych punktów dostępu pochodzących ze słownika jakiegoś języka informacyjno-wyszukiwawczego[16]. Wreszcie zewnętrzna autogeneracja metadanych zakłada ekstrakcję metadanych niezawartych wewnątrz dokumentu[17]. Należy podkreślić, że obecnie mamy do czynienia w dużym stopniu z narzędziami, które wymagają pewnego nadzoru ze strony obsługujących je ludzi. W takim przypadku można mówić wyłącznie o opracowaniu semi-automatycznym[18]. Szczególnie interesujące z punktu widzenia bibliotek wydają się być dwie z wymienionych wyżej metod: ekstrakcja treści oraz automatyczne indeksowanie z tym, że obie te metody wymagają dostępu do cyfrowych tekstów opracowywanych publikacji.

Tradycyjnie opracowanie zbiorów w bibliotekach dzieli się na dwa podstawowe rodzaje, które wymagają nieco innych predyspozycji i umiejętności:

- opracowanie formalne;
- opracowanie rzeczowe.

Oba rodzaje opracowania różnią się znacznie, jeśli chodzi o wymagania stawiane przed komputerem. Komputery przydają się szczególnie w pracach, w których liczy się szybkość, a poprawność lub niepoprawność wykonania zadania jest łatwa do określenia[19]. Należy jednocześnie pamiętać, że komputery coraz lepiej radzą sobie również ze skomplikowanymi, nierutynowymi zadaniami. Opracowanie formalne polega w dużym stopniu na wyodrębnianiu informacji z dokumentu i zapisywaniu ich w odpowiednich polach rekordu bibliograficznego, co oznacza, że komputery przy zapewnieniu odpowiednich warunków są w stanie dobrze wykonywać taką pracę. Opracowanie rzeczowe jest bardziej złożonym zadaniem, jednak komputery również i z nim radzą sobie coraz lepiej. Poniżej zaprezentowano przykłady projektów wykorzystujących ekstrakcję treści i automatyczne indeksowanie zarówno w World Wide Web (WWW), jak i w bibliotekach.

W 1997 r. powstała pierwsza internetowa wyszukiwarka treści naukowych – CiteSeer, która dziewięć lat później została przemianowana na CiteSeerX, pod którą to nazwą funkcjonuje do dnia dzisiejszego. José Luis

[14] J. Park, A. Brenza, *op.cit.*, s. 25.

[15] *Ibidem*, s. 29.

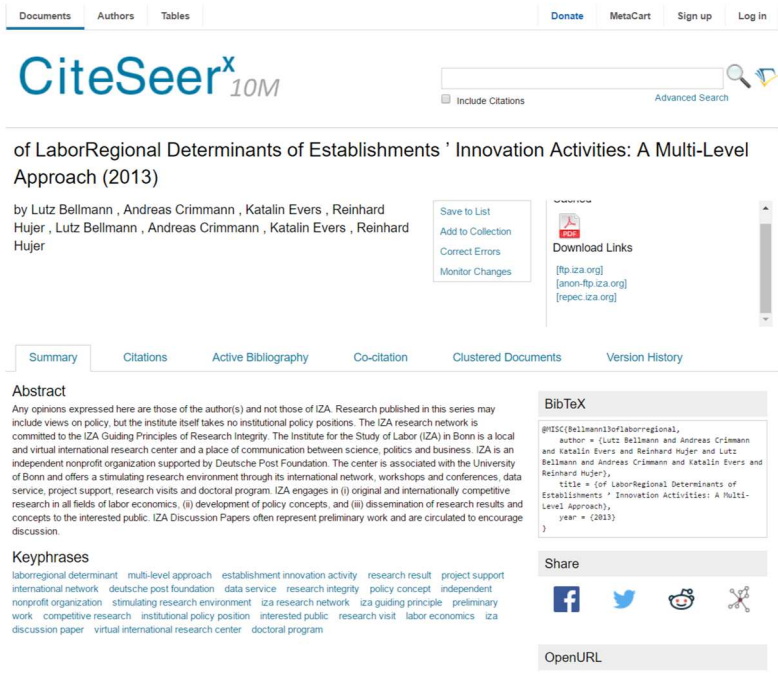
[16] *Ibidem*, s. 32.

[17] *Ibidem*, s. 35.

[18] *Ibidem*, s. 22–23.

[19] W. Randtke, *Automated Metadata Creation: Possibilities and Pitfalls*, „The Serials Librarian" 2013, vol. 64, issue 1–4, s. 273.

Ortega podkreślił, że istotnym wkładem tej wyszukiwarki w rozwój internetowych narzędzi naukowych było stworzenie autonomicznego indeksu cytowań opartego na automatycznej analizie plików tekstowych zawierających treść artykułów[20]. Podobnie w automatyczny sposób były wyodrębniane elementy opisu bibliograficznego takie jak autorzy, tytuły, a ponadto abstrakty. Zadanie jest trudne, ponieważ ekstrakcja metadanych następuje ze zróżnicowanych strukturalnie dokumentów. Do wyodrębnienia cytatów wykorzystywano pakiet ParsCit, a do pozostałych danych – SVM HeaderParse. Podstawowy problem, podkreślony przez Ortegę, polegał na błędnym wyodrębnianiu danych z pełnych tekstów. Dodatkowo brakowało jakiejkolwiek standaryzacji między innymi w nazwach autorów oraz tytułach czasopism[21]. Przykład takiego błędu w wyodrębnianiu danych pokazano na rysunku (rys. 1).



Rys. 1. Przykład niepoprawnego wyodrębnienia tytułu w wyszukiwarce CiteSeerX

Źródło: *CiteSeerX*, [dostęp: 6.04.2017], http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.377.2834&rank=2.

[20] J.L. Ortega, *Academic Search Engines: A Quantitative Outlook*, Amsterdam 2014, s. 12.
[21] *Ibidem*, s. 20–24.

Dane dotyczące dokumentów z wyszukiwarki CiteSeerX, opublikowane ostatnio przez badaczy pracujących nad jej rozwojem, są według nich unikalne w stosunku do innych tego rodzaju danych pod względem ujednoznacznienia nazw autorskich w głównej bazie. W celu poprawy jakości wyodrębniania metadanych badacze zastąpili pakiet SVM HeaderParse innym narzędziem osiągającym lepsze wyniki – GROBID[22].

Jednak w WWW istnieją narzędzia, które dobrze radzą sobie z wyodrębnianiem metadanych. Przykładem jest wyszukiwarka naukowa Google Scholar. Ortega podkreślił, że poprawienie autonomicznego indeksu cytowań oraz zastosowanie robotów internetowych Google pozwoliło stworzyć największą, dostępną za darmo bazę treści naukowych w Internecie[23]. Według stron pomocy serwisu, aby zapewnić możliwość indeksowania treści, należy zapisać metadane w odpowiednich metatagach HTML lub odpowiednio sformatować dokument, na przykład tytuł powinien być zapisany największą czcionką u góry strony, a nazwy autorów nieco mniejszą czcionką poniżej lub powyżej tytułu[24]. Odpowiednie sformatowanie dokumentu ułatwia ekstrakcję treści i uniknięcie podczas tego procesu błędnego przyporządkowania metadanych[25].

W bibliotekach również prowadzone są projekty oparte na ekstrakcji informacji z dokumentów. Jeden z nich miał na celu digitalizację i opracowanie zbioru przepisów administracyjnych stanu Floryda – „Florida Administrative Code" (FAC). W latach 1963–1983 FAC był wydawany w postaci skoroszytowego wydawnictwa wymiennokartkowego. Uzupełnieniem były suplementy zawierające instrukcje na temat dodawania, zamiany lub usuwania konkretnych stron[26]. Poszczególne strony miały stały układ, co pozwoliło na zastosowanie automatycznego wyodrębnienia danych. W ten sposób uzupełniono 99,3% numerów rozdziałów przed myślnikiem[27], 92,2% numerów rozdziałów

[22] J. Wu [et al.], *CiteSeerX Data: Semanticizing Scholarly Papers*, [w:]: *Proceedings of the International Workshop on Semantic Big Data*, New York 2016, s. 2–4, https://doi.org/10.1145/2928294.2928306, [dostęp: 6.04.2017].

[23] J.L. Ortega, *op. cit.*, s. 138.

[24] *Inclusion Guidelines for Webmasters: Indexing Guidelines*, https://scholar.google.com/intl/en/scholar/inclusion.html#indexing, [dostęp: 6.04.2017].

[25] Mechanizm ekstrakcji metadanych z tekstu jest również stosowany między innymi w Mendeleyu, programie do zarządzania danymi bibliograficznymi. Na stronie WWW oprogramowania wskazano, że jakość wyodrębnionych metadanych zależy między innymi od złożoności układu artykułu (*Mendeley | How Does the Automatic Document Details Extraction Work*, http://support.mendeley.com/customer/en/portal/articles/227883-how-does-the-automatic-document-details-extraction-work-), [dostęp: 6.04.2017].

[26] W. Randtke, *op. cit.*, s. 267.

[27] Każdy numer rozdziału składał się z dwóch liczb rozdzielonych myślnikiem (W. Randtke, *op. cit.*, s. 272).

po myślniku, 93,6% numerów stron oraz 88,4% numerów suplementów[28]. Brakujące wartości zostały uzupełnione przez specjalnie zatrudnionych w tym celu studentów. W celu porównania liczby błędów popełnionych przez ludzi i przez komputer zestawiono odpowiednie arkusze kalkulacyjne. Poziom błędów był zróżnicowany. Dla numeru suplementu wynosił 0,8% (ludzie) i 2,4% (komputer), a dla numerów stron 3,1% (ludzie) i 1% (komputer). Podsumowując wyniki projektu, Wilhelmina Randtke stwierdziła, że poziom błędów był porównywalny, chociaż rozpowszechnione jest mniemanie, że metadane tworzone komputerowo są niższej jakości niż te tworzone manualnie. Zastosowanie komputerów pozwoliło znacznie zwiększyć ilość wprowadzonych metadanych[29]. Podsumowując należy podkreślić, że sukces automatycznego opracowania formalnego w dużym stopniu zależy od powtarzalności rozmieszczenia wyodrębnianych danych w dokumentach.

Automatyczne opracowanie rzeczowe jest dużo bardziej złożonym problemem niż opracowanie formalne. Jedną z takich prób podjęło przedsiębiorstwo Microsoft w nowej odsłonie swojej wyszukiwarki naukowej nazwanej Microsoft Academic (MA). Na głównej stronie tego serwisu znajduje się ułożony hierarchicznie zbiór dziedzin i tematów (fields of study), według których można przeglądać zindeksowane publikacje naukowe. Nazwy dziedzin i tematów mogą być również wykorzystywane w głównym oknie wyszukiwarki[30]. Manlio De Domenico, Elisa Omodei oraz Alex Arenas zwrócili uwagę na niejasny mechanizm przypisywania dziedzin i tematów do konkretnych artykułów. Podkreślili, że znaleźli wiele błędnie opisanych prac naukowych, jak na przykład artykuł z dziedziny agronomii, do którego przyporządkowano między innymi temat „Ogólna teoria względności" (general relativity)[31]. Przegląd opisów w serwisie potwierdza spostrzeżenia cytowanych autorów. Na przykład artykuł autorstwa Bartłomieja Włodarczyka zatytułowany „Mapy tematów jako system reprezentacji wiedzy" otrzymał następujące tematy: „przetwarzanie języka naturalnego" (natural language processing), „językoznawstwo" (linguistics), „rozpoznawanie mowy" (speech recognition). Źródłem opisu były metadane

[28] W. Randtke, *op. cit.*, s. 279.

[29] *Ibidem*, s. 281–283.

[30] Badania prowadzone obecnie w Microsoft Research zmierzają do usprawnienia wyszukiwania zasobów naukowych. Zob.: A. Sinha [et al.], *An Overview of Microsoft Academic Service (MAS) and Applications*, [w:] *Proceedings of the 24th International Conference on World Wide Web Companion (WWW 2015 Companion)*, New York 2015, s. 243–246, https://doi.org/10.1145/2740908.2742839, [dostęp: 6.04.2017].

[31] M. De Domenico, E. Omodei, A. Arenas, *Quantifying the Diaspora of Knowledge in the Last Century*, s. 2, http://arxiv.org/abs/1604.00696, [dostęp: 6.04.2017].

z serwisu Biblioteka Nauki Centrum Otwartej Nauki. Oprócz krótkiego abstraktu w języku polskim i angielskim znajdują się tam również angielski wariant tytułu i dwa słowa kluczowe w tym języku: „topic maps", „subject headings' language"[32]. Wynikiem automatycznej analizy były wyrażenia nie mające nic wspólnego z tekstem artykułu, takie jak na przykład „rozpoznawanie mowy". Dodatkowo Sven E. Hug, Michael Ochsner oraz Martin P. Braändle pokazali niespójność hierarchii tematów w MA. Termin „Nauki społeczne" występuje na drugim poziomie hierarchii i jest podrzędny do tematów „Psychologia" oraz „Socjologia". Na tym samym poziomie, co nauki społeczne, znajduje się także wąski zakresowo temat „Cykl sonatowy"[33]. Problemy występują więc zarówno na etapie projektowania systemu, jak i automatycznego indeksowania zasobów naukowych. Trzeba jednak podkreślić, że zadanie jest bardzo trudne, biorąc pod uwagę szeroki zakres dziedzin uwzględnionych w MA.

Projekty dotyczące automatycznego indeksowania rzeczowego są rozwijane również w bibliotekach. W 2009 r. w Niemieckiej Bibliotece Narodowej (NBN) podjęto decyzję o zaprzestaniu od kolejnego roku manualnego opracowania monografii elektronicznych. W latach 2009–2011 oraz w 2012 i 2013 r. przeprowadzono projekt Petrus, którego celem było sprawdzenie możliwości wprowadzenia takiej zmiany. Przyczyną kontynuacji opracowania rzeczowego dokumentów elektronicznych była między innymi chęć zapewnienia spójności danych. Wybrano niemieckie oprogramowanie Averbis Extraction Platform opracowane przez przedsiębiorstwo Averbis specjalizujące się w automatycznym indeksowaniu dokumentów medycznych[34]. Przeprowadzone w NBN testy objęły dokumenty z 12 dziedzin wiedzy, a więc były zróżnicowane pod względem treściowym. Poziom kompletności wyniósł między 0,5 a 0,9, jednak dokładność była na niewystarczającym poziomie. Według Ulriki Junger, autorki artykułu podsumowującego projekt, oprogramowanie przydzielało za dużo niepoprawnych haseł, przypisując jednocześnie za mało użytecznych. Podkreśliła ona także, że nie osiągnięto jeszcze etapu wdrożenia[35].

[32] *Mapy tematów jako system reprezentacji wiedzy*, http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.cejsh-b170164a-93d9-4324-80d9-d99351d718dc, [dostęp: 6.04.2017].

[33] S.E. Hug, M. Ochsner, M.P. Brändle, *Citation Analysis with Microsoft Academic*, „Scientometrics" 2017, vol. 111, issue 1, s. 374.

[34] U. Junger, *Can Indexing Be Automated? The Example of the Deutsche Nationalbibliothek*, „Cataloging & Classification Quarterly" 2014, vol. 52, issue 1, s. 104–105.

[35] *Ibidem*, s. 107.

Inne przedsięwzięcia dotyczą konkretnych dziedzin lub dyscyplin naukowych. Jeden z najbardziej udanych projektów z zakresu semi-automatycznego indeksowania rzeczowego jest rozwijany od 15 lat w U.S. National Library of Medicine (NLM). Biblioteka ta prowadzi bazę bibliograficzną MEDLINE zawierającą opisy artykułów z dziedziny nauk biologicznych i medycznych[36]. MEDLINE wchodzi w skład bazy PubMed, w której znajdują się również opisy artykułów z innych dziedzin. NLM utrzymuje także repozytorium PubMed Central zawierające pełne teksty artykułów medycznych i biologicznych. W 1996 r. powołano w NLM zespół pod nazwą Indexing Initiative, którego celem była, w związku z rosnącą liczbą dokumentów do opracowania i zmniejszającymi się zasobami, analiza nowych możliwości związanych z indeksowaniem artykułów wchodzących w skład bazy MEDLINE[37]. Skala zadania stojącego przed osobami zajmującymi się opracowaniem tych zasobów jest bardzo duża. Przykładowo w roku fiskalnym 2016 zindeksowano aż 869 666 artykułów[38]. Głównym osiągnięciem opisywanej grupy badawczej było zaprojektowanie i uruchomienie w 2002 r. narzędzia do semi-automatycznego opracowania zbiorów biomedycznych nazwanego Medical Text Indexer (MTI). Oprogramowanie to poddawane jest stałym ulepszeniom tak, aby zwiększyć dokładność i kompletność przydzielanych terminów oraz sprostać oczekiwaniom osób indeksujących artykuły medyczne. Obecnie podstawą przydzielania haseł MeSH są tytuły oraz abstrakty publikacji, jednak zespół pracuje nad możliwościami wykorzystania pełnych tekstów artykułów. Podstawowym zadaniem MTI jest dostarczenie zbioru rekomendacji, na które składają się deskryptory MeSH, modyfikatory oraz check tagi. Każdy opracowujący może, ale nie musi, korzystać z tych podpowiedzi. Dane zaprezentowane przez twórców oprogramowania pokazują stały wzrost użycia narzędzia. W 2002 r. indeksatorzy odwoływali się do niego w przypadku 15,75% opracowywanych artykułów, a w 2014 r. w przypadku 62,44% dokumentów[39]. Oprócz tego podstawowego zastosowania MTI w NLM w 2011 r. zdecydowano, że niektóre czasopisma, w przypadku których narzędzie osiągało najlepsze rezultaty, jeśli chodzi o dokładność i kompletność opracowania, będą w pierwszej fazie indeksowane wyłącznie przez to oprogramowanie bez ingerencji człowieka. Dopiero w kolejnej fazie korekty doświadczeni opracowujący sprawdzają opis dodając

[36] *Fact Sheet MEDLINE, PubMed and PMC (PubMed Central): How are they different*, https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html, [dostęp: 6.04.2017].

[37] J. Mork, A. Aronson, D. Demner-Fushman, *12 Years on – Is the NLM Medical Text Indexer Still Useful and Relevant?*, „Journal of Biomedical Semantics" 2017, vol. 8, s. 2.

[38] *Key MEDLINE Indicators*, https://www.nlm.nih.gov/bsd/bsd_key.html, [dostęp: 6.04. 2017].

[39] J. Mork, A. Aronson, D. Demner-Fushman, *op. cit.*, s. 5.

niezbędne hasła i usuwając błędnie przydzielone przez komputer. Tę ścieżkę opracowania nazwano MTI First Line (MTIFL)[40]. 26 maja 2016 r. zbiór ten liczył 489 czasopism[41]. Ponadto dla 51 czasopism, które wypadały szczególnie dobrze w testach dokładności i kompletności, wprowadzono dodatkowe filtrowanie skoncentrowane na podniesieniu tego pierwszego parametru, co znacznie poprawiło jakość opracowania. Poniżej przedstawiono tabelę zawierającą wartości współczynnika F_1[42] dla różnych ścieżek opracowania w latach 2007 i 2015 (tab. 2).

Tabela 2

Współczynnik F_1 dla różnych ścieżek opracowania w latach 2007 i 2015

	2007	2015
MTI	0.3810	0.5878
MTIFL	-	0.7113
Zbiór 51 wybranych czasopism	-	0.8642

Źródło: na podstawie D. Demner-Fushman, J. Mork, *NLM Medical Text Indexer Technical Report to the LHNCBC Board of Scientific Counselors April 2016*, Bethesda 2016, s. 14, https://lhncbc.nlm.nih.gov/system/files/pub9359.pdf, [dostęp: 6.04.2017].

Widoczna jest znaczna poprawa współczynnika F_1 pomiędzy 2007 a 2015 r. o 0,2068, co wskazuje na pozytywny rozwój oprogramowania. Zwracają uwagę także znaczne różnice między ścieżkami opracowania, na przykład w przypadku 51 wybranych czasopism różnica w stosunku do podstawowej wersji MTI wynosi 0,2764[43]. Pokazuje to, że oprogramowanie cały czas ma szansę na uzyskiwanie lepszych rezultatów w stosunku do większej liczby czasopism.

Należy podkreślić, że MTI powstaje w ścisłej współpracy z osobami opracowującymi artykuły do bazy MEDLINE[44]. Ponadto zespół tworzący oprogramowanie opracował razem z Sekcją MeSH uproszczony interfejs użytkownika umożliwiający analizę dowolnego tekstu z poziomu specjalnej strony

[40] *Ibidem*, s. 2–3.

[41] *MTI First Line (MTIFL) Indexing*, s. 1, https://ii.nlm.nih.gov/MTIMTIFL_Journal_List.pdf, [dostęp: 6.04.2017].

[42] Współczynnik F_1 to średnia harmoniczna dokładności i kompletności.

[43] D. Demner-Fushman, J. Mork, *op. cit.*, s. 14.

[44] J. Mork, A. Aronson, D. Demner-Fushman, *op. cit.*, s. 9.

internetowej[45]. Oprócz prowadzenia wewnętrznej ewaluacji, której wyniki podano powyżej, zespół MTI bierze udział w konkursach BioASQ Challenge. Umożliwiają one porównanie wydajności różnych systemów tworzonych z myślą o semantycznym indeksowaniu zasobów biomedycznych. Jak podkreślają twórcy oprogramowania, spotkania te nie tylko są okazją do porównania programów, ale również stanowią forum wymiany pomysłów pomagających w ulepszeniu MTI[46]. Biorąc pod uwagę zakres współpracy z podmiotami wewnętrznymi i zewnętrznymi, czas trwania projektu oraz osiągane wyniki należy stwierdzić, że MTI może stanowić przykład wzorcowego rozwoju tego typu oprogramowania.

Stosunek bibliotekarzy do automatycznego indeksowania zbiorów

Istotną kwestią w przypadku wprowadzania każdego nowego rozwiązania jest jego akceptacja przez osoby, które będą z niego w przyszłości korzystać. W tej części artykułu porównano wyniki badań dotyczących opinii bibliotekarzy niemieckojęzycznych, anglojęzycznych oraz polskich na temat automatycznego indeksowania zbiorów.

W artykule na temat zmian w zasadach katalogowania Sally Glasser napisała: „W celu sprostania wymaganiom dzisiejszych stanowisk związanych z katalogowaniem studenci bibliotekoznawstwa muszą zdobyć szerszy zbiór umiejętności, które oprócz tradycyjnej teorii i praktyki zasad rejestracji bibliograficznej oraz standardów metadanych obejmują także umiejętności w zakresie zarządzania, obsługi komputera, umiejętności komunikacyjne, pracy w zespole, elastyczność, i być może najważniejszą, gotowość do uczenia się i zdolność radzenia sobie ze zmianami”[47]. Stwierdzenie to można z powodzeniem odnieść nie tylko do studentów, ale również do wszystkich bibliotekarzy pracujących obecnie w bibliotekach. Przed opisaniem wyników badań dotyczących opinii specjalistów od opracowania na temat automatycznego indeksowania, warto zastanowić się nad dostępnymi dla polskich bibliotekarzy źródłami wiedzy na ten temat. Pomijając opracowania informatyczne, jedną z niewielu

[45] Vide *MeSH on Demand*, https://www.nlm.nih.gov/mesh/MeSHonDemand.html, [dostęp: 6.04.2017].

[46] J. Mork, A. Aronson, D. Demner-Fushman, *op. cit.*, s. 7–8. Opis jednego z nowszych ulepszeń inspirowanych spotkaniami BioASQ vide I. Zavorin, J.G. Mork, D. Demner-Fushman, *Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results*, [w:] *Proceedings of the Fourth BioASQ Workshop*, s. 8–15, http://aclweb.org/anthology/W/W16/W16-3102.pdf, [dostęp: 06.04.2017].

[47] S. Glasser, *The Changing Face of Cataloging Positions at Academic Institutions: What Skill Set is Needed, and How Can Students Prepare?*, „The Serials Librarian” 2007, vol. 51, issue 3–4, s. 48.

polskich publikacji napisanych z zakresu bibliologii i informatologii, a poświęconych tej tematyce jest książka Piotra Malaka[48]. Istotnym zagadnieniem wydaje się dostępność treści z zakresu automatycznego generowania metadanych w programach nauczania studiów związanych ze wspomnianymi dziedzinami.

W celu ustalenia czy są one uwzględniane na polskich uczelniach przejrzano programy nauczania i plany zajęć polskich uczelni oferujących studia przygotowujące do pracy w bibliotece. Są one dostępne na stronach WWW tych instytucji[49]. Nie znaleziono w nich bezpośredniego odwołania do tej problematyki. Jedynie w sylabusie przedmiotu „Organizacja informacji i wiedzy" prowadzonym w ramach kierunku „Zarządzanie informacją" na Uniwersytecie Jagiellońskim wśród zalecanych lektur znaleziono książkę autorstwa Malaka. Nie znaczy to, że treści z omawianego zakresu nie są w ogóle uwzględnione w toku nauczania przyszłych bibliotekarzy i informatologów. Prowadzący mogą je uwzględniać w jakimś stopniu podczas prowadzenia zajęć, ważne jest jednak to, że nie są one podkreślane w programach nauczania i podczas prowadzonych zajęć. Zwraca również uwagę brak szkoleń w tym zakresie. W broszurze reklamowej, prezentującej jedno z niewielu takich szkoleń przeznaczonych dla bibliotekarzy zapisano, że: „Celem warsztatów jest zaprezentowanie technik przetwarzania elektronicznych dokumentów tekstowych na potrzeby automatycznego indeksowania oraz wyszukiwania informacji"[50]. W tym kontekście warto wspomnieć o badaniu przeprowadzonym przez Jung-rana Parka i Yuji Tosakę, którego celem była ocena stanu kształcenia ustawicznego amerykańskich bibliotekarzy zajmujących się katalogowaniem i metadanymi. Jedno z pytań zadanych badanym podczas ankiety internetowej dotyczyło tematyki szkoleń, w których ostatnio uczestniczyli.

Wśród najczęściej wymienianych tematów znalazły się Resource Description and Access (80,4%), zaawansowane katalogowanie (53,1%), Functional Requirements for Bibliographic Records (45,3%) i standardy z zakresu

[48] Zob.: P. Malak, *Indeksowanie treści: porównanie skuteczności metod tradycyjnych i automatycznych*, Warszawa 2012. Analizę bibliometryczną publikacji dotyczących automatycznego indeksowania z lat 1956–2000 zawiera artykuł Antonio Pulgarína oraz Isidoro Gil-Leivy. Vide A. Pulgarín, I. Gil-Leiva, *Bibliometric Analysis of the Automatic Indexing Literature: 1956–2000*, „Information Processing & Management" 2004, vol. 40, issue 2, s. 365–377.

[49] Podczas przeglądu uwzględniono następujące uczelnie: Uniwersytet Kazimierza Wielkiego w Bydgoszczy, Uniwersytet w Białymstoku, Uniwersytet Jagielloński, Uniwersytet Łódzki, Uniwersytet Marii Curie-Skłodowskiej w Lublinie, Uniwersytet Mikołaja Kopernika w Toruniu, Uniwersytet Wrocławski, Uniwersytet im. Adama Mickiewicza w Poznaniu, Uniwersytet Śląski w Katowicach oraz Uniwersytet Warszawski.

[50] *Indeksowanie treści w teorii i praktyce: warsztaty*, s. [2], http://www.novaskills.pl/docs/Indeksowanie_tresci.pdf, [dostęp: 6.04.2017].

kontroli autorytatywnej oraz słownictwa kontrolowanego (39,6%). Wśród odpowiedzi pojawiły się również takie, w których jako temat szkoleń wymieniono narzędzia do semi-automatycznego generowania metadanych. Uczestniczyło w nich jednak niewielu bibliotekarzy (3,7%), podobnie jak na przykład w tych dotyczących ontologii (6,6%), czy też standardu Simple Knowledge Organization System (2,9%)[51]. (Semi-)automatyczne indeksowanie nie cieszy się więc zbyt dużym zainteresowaniem mierzonym liczbą przeprowadzonych szkoleń. Przeważają szkolenia ściśle związane z bieżącą praktyką biblioteczną. W tym kontekście należy zadać pytanie, jak to ograniczone zainteresowanie przełoży się na przyszłość bibliotek jako centrów dostępu do informacji, czy bez tych umiejętności biblioteki mogą pozostać ważnym pośrednikiem w dostępie do różnego rodzaju dokumentów. Trudno jest udzielić zdecydowanej odpowiedzi, jednak wydaje się, że w kontekście wzrastającej liczby zasobów i wobec szybko postępującej komputeryzacji, umiejętności te mogą okazać się istotne.

Badanie przeprowadzone przez Alice Keller wśród niemieckojęzycznych i anglojęzycznych bibliotekarzy miało na celu zbadanie ich postaw wobec opracowania rzeczowego, ze szczególnym uwzględnieniem indeksowania automatycznego. Między badanymi grupami występuje różnica w sposobie organizacji pracy. W Niemczech opracowanie formalne jest oddzielone od rzeczowego i zajmują się nimi odrębne grupy pracowników, podczas gdy w krajach anglojęzycznych jedna osoba jest zazwyczaj odpowiedzialna za całość opracowania dokumentu. W związku z tym Keller postanowiła zbadać również, czy wspomniana różnica w organizacji prac katalogowych ma wpływ na opinie wyrażane przez bibliotekarzy. Internetową ankietę wypełniło ostatecznie 114 bibliotekarzy niemieckojęzycznych (N) i 61 anglojęzycznych (A) pracujących w bibliotekach akademickich i narodowych[52]. W niniejszym artykule skupiono się głównie na odpowiedziach dotyczących automatycznego indeksowania zbiorów[53]. Przeprowadzono też wstępne badanie opinii pracowników polskich bibliotek akademickich na temat komputeryzacji opracowania. Uzyskane odpowiedzi zostały następnie porównane z wynikami badania Keller.

[51] J. Park, Y. Tosaka, *Advancing Professional Learning in Libraries: An Exploratory Study of Cataloging and Metadata Professionals' Experiences and Perspectives on Continuing Education Issues*, „Cataloging & Classification Quarterly" 2017, vol. 55, issue 3, s. 162–163.

[52] A. Keller, *Attitudes among German- and English-Speaking Librarians toward (Automatic) Subject Indexing*, „Cataloging & Classification Quarterly" 2015, vol. 53, issue 8, s. 895–896.

[53] Pominięto kwestię oceny współkatalogowania jako sposobu na przyśpieszenie tempa opracowania zasobów.

Internetową ankietę rozesłano w marcu 2017 r., z wykorzystaniem serwisu Formularze Google[54], do bibliotekarzy zajmujących się opracowaniem rzeczowym zbiorów w bibliotekach uniwersyteckich oraz politechnicznych. Udział w badaniu był dobrowolny, a kwestionariusz nie zawierał pytań umożliwiających identyfikację osób oraz bibliotek. Zapewnienie anonimowości miało na celu zachęcenie bibliotekarzy do wypełnienia ankiety, zarazem jednak ograniczyło możliwość uzyskania szczegółowych odpowiedzi. Kwestionariusz wypełniło 111 bibliotekarzy: 75 z bibliotek uniwersyteckich oraz 36 pracujących w bibliotekach politechnik. W artykule zaprezentowano odpowiedzi na 10 pytań wiążących się z indeksowaniem automatycznym. Pytania te były wzorowane na zadanych przez Keller tak, aby możliwe było porównanie opinii bibliotekarzy zagranicznych i polskich. Ograniczono się przy tym do indeksowania dokumentów tekstowych, pomijając kwestię automatyzacji opracowania dokumentów ikonograficznych.

Większość polskich bibliotekarzy, którzy wzięli udział w ankiecie, zajmuje się zarówno opracowaniem rzeczowym, jak i formalnym książek lub artykułów (75,5%), co jest porównywalne z odpowiedziami uzyskanymi od bibliotekarzy anglojęzycznych. Większość bibliotekarzy w badaniu Keller określiło swoje umiejętności w zakresie opracowania rzeczowego jako zaawansowane (N – 66,7%; A – 60,7%)[55]. Odpowiedzi polskich respondentów również były podobne – 62,2%[56]. We wszystkich grupach przeważały więc osoby o wysokiej znajomości problematyki indeksowania rzeczowego. Keller podkreśliła, że proces wprowadzania zmian kończy się sukcesem, jeśli uczestniczące w nim osoby rozumieją i akceptują konieczność jego przeprowadzenia[57]. W związku z tym w kwestionariuszu pojawiły się pytania o czynniki mogące potencjalnie wpływać na zwiększenie akceptacji indeksowania automatycznego. Na pytanie o redukcję liczby pracowników zajmujących się opracowaniem rzeczowym w ciągu ostatnich 5 lat większość ankietowanych w badaniu Keller odpowiedziała, że nastąpiło niewielkie zmniejszenie ich liczby (N – 36,8%; A – 47,5%) lub też poziom zatrudnienia nie zmienił się (N – 30,7%; A – 36,1%)[58]. Nieco odmienne wyniki uzyskano w przypadku polskich respondentów – 45% z nich odpowiedziało, że liczba pracowników nie zmieniła się,

[54] Vide *Formularze Google*, https://www.google.pl/intl/pl/forms/about/, [dostęp: 6.04. 2017].

[55] A. Keller, *op. cit.*, s. 897.

[56] W badaniu określono, że osoba doświadczona to taka, która zajmuje się opracowaniem rzeczowym powyżej 5 lat.

[57] A. Keller, *op. cit.*, s. 903.

[58] *Ibidem*, s. 898.

a 28%, że się zwiększyła[59]. W badaniu przeprowadzonym wśród zagranicznych bibliotekarzy zapytano również o to, czy występują niedobory zasobów ludzkich, jeśli chodzi o opracowanie rzeczowe zbiorów. Większość ankietowanych odpowiedziała, że występują czasami w niektórych dziedzinach (N – 41,2%; A – 45,9%)[60]. W kwestionariuszu rozesłanym do polskich bibliotekarzy znalazło się identyczne pytanie, jednak z innymi odpowiedziami (tak, nie, nie wiem). Zdecydowana większość ankietowanych odpowiedziała, że w ich bibliotekach nie występują takie problemy (67,6%). Uzyskane odpowiedzi nie wskazują na zapotrzebowanie na wprowadzenie indeksowania automatycznego. Kwestionariusz przygotowany przez szwajcarską badaczkę zawierał pytanie o prawdziwość czterech stwierdzeń, do których każdy respondent odnosił się, korzystając z czterostopniowej skali Likerta (1 – nieprawda; 4 – zdecydowanie prawda). Jedno z nich brzmiało: „Musimy zupełnie od nowa zastanowić się nad opracowaniem rzeczowym. Obecne modele i zasady są całkowicie nieaktualne”[61]. Stwierdzenie to zyskało umiarkowane poparcie zarówno wśród bibliotekarzy niemieckojęzycznych (2,80), jak i anglojęzycznych (2,36). Większa potrzeba zmian jest jednak zauważana wśród tej pierwszej grupy bibliotekarzy[62]. Pytanie zadane polskim bibliotekarzom różniło się od zadanego w badaniu Keller. Zapytano o to, czy opracowanie rzeczowe dokumentów wymaga znaczących zmian, które ułatwią pracę bibliotekarzom, a użytkownikom wyszukiwanie. Możliwe były trzy odpowiedzi (tak, nie, nie wiem). Większość ankietowanych odpowiedziała na to pytanie twierdząco (53,2%), a więc wśród badanych bibliotekarzy istnieje świadomość konieczności wprowadzenia zmian w opracowaniu zbiorów.

Trzy pytania z kwestionariusza Keller odnosiły się bezpośrednio do indeksowania automatycznego. Pierwsze z nich dotyczyło uczestnictwa w programach i projektach z tego zakresu. Zdecydowana większość ankietowanych nie brała udziału w takich przedsięwzięciach (N – 70,2%; A – 82%). Takie projekty są o wiele częstsze w bibliotekach niemieckiego obszaru językowego (22,8%), niż w bibliotekach krajów anglojęzycznych (3,2%)[63]. W ankiecie skierowanej do polskich bibliotekarzy zdecydowano się zadać dwa pytania

[59] Odpowiedzi nie dotyczą sytuacji bibliotek, lecz osób, które wzięły udział w ankiecie. W celu zbadania zmian w zatrudnieniu w bibliotekach należałoby przejrzeć sprawozdania roczne z działalności tych instytucji, ewentualnie przeprowadzić ankietę wśród dyrektorów tych placówek.

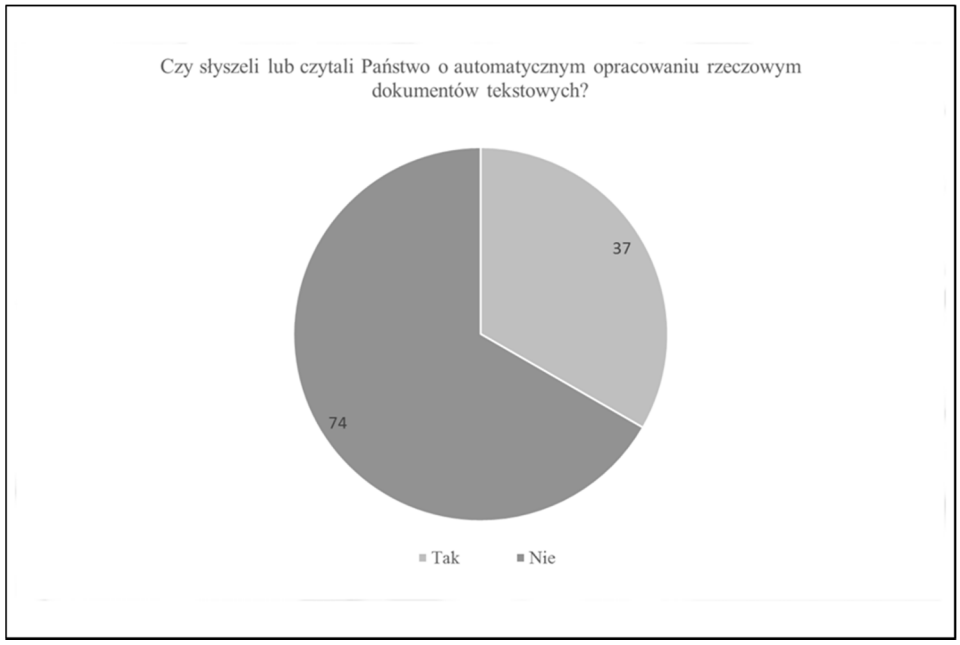
[60] A. Keller, *op. cit.*, s. 898.

[61] *Ibidem*, s. 902.

[62] *Ibidem*.

[63] *Ibidem*, s. 899–900. Keller wyjaśnia w przypisie, że część podanych przez bibliotekarzy projektów należałoby zaliczyć raczej do programów wzbogacania danych, a więc podane liczby powinny być w rzeczywistości mniejsze (A. Keller, *op. cit.*, s. 904).

ze względu na przypuszczenie, że niewielu z nich miało możliwość udziału w tego rodzaju programach. Pierwsze z nich brzmiało: „Czy słyszeli lub czytali Państwo o automatycznym opracowaniu rzeczowym dokumentów tekstowych?" (rys. 2).



Rys. 2. Znajomość problematyki automatycznego opracowania rzeczowego dokumentów tekstowych wśród polskich bibliotekarzy pracujących w bibliotekach akademickich
Źródło: opracowanie własne, marzec 2017 r.

Większość respondentów odpowiedziała na nie negatywnie (66,7%). Pokazuje to niską świadomość polskich bibliotekarzy w zakresie indeksowania automatycznego. Warto przyjrzeć się pozytywnym odpowiedziom z podziałem na pracowników bibliotek uniwersyteckich i politechnik. Spośród pracowników politechnik, którzy wzięli udział w ankiecie, 47,2% słyszało lub czytało o automatycznym indeksowaniu, podczas gdy w przypadku bibliotekarzy z uniwersytetów odsetek ten był znacznie niższy – 26,7%. Wyniki wskazują na stosunkowo niewielką wiedzę na temat tego sposobu opracowania zbiorów, przy czym lepszą znajomością tej problematyki charakteryzują się bibliotekarze pracujący na uczelniach technicznych. W związku z wyspecjalizowaną kadrą właśnie na politechnikach takie projekty wydają się mieć większą szansę na realizację[64].

Kolejne pytanie skierowane do polskich bibliotekarzy dotyczyło ich udziału w konkretnym projekcie. Spośród przebadanych nieco ponad 97% nie

miało okazji uczestniczenia w takim przedsięwzięciu. Odpowiedź pozytywną zaznaczyły tylko trzy osoby (2,7%), w tym dwie z politechniki. W związku z chęcią zachowania anonimowości nie zapytano o konkretny projekt. Można powiedzieć, że projekty z zakresu automatycznego indeksowania są niemal nieobecne w przebadanych bibliotekach angielskiego obszaru językowego i polskich. Następne pytanie dotyczyło możliwości zastąpienia opracowania manualnego przez indeksowanie automatyczne. Jego celem było również poznanie opinii respondentów jak szybko może to nastąpić. W przypadku bibliotekarzy niemieckiego kręgu językowego najczęściej wybieraną odpowiedzią była ta, że nie nastąpi to nigdy (40,4%), a wśród bibliotekarzy anglojęzycznych odpowiedź – nie wiem (29,5%). Kolejne pod względem liczby zaznaczeń odpowiedzi dotyczyły perspektywy czasowej tych zmian, czyli czy stanie się to do 2020 (N – 13,2%; A – 11,5%) czy też do 2025 r. (N – 17,5%; A – 16,4%). Ogólnie należy podkreślić, że bibliotekarze z krajów niemieckojęzycznych byli nieco bardziej pewni możliwości szybszej automatyzacji opracowania niż bibliotekarze anglojęzyczni[65]. Na identycznie sformułowane pytanie polscy bibliotekarze odpowiedzieli odmiennie. Największy odsetek stwierdził, podobnie jak w krajach niemieckojęzycznych, że taka zmiana w ogóle nie nastąpi (48,7%). Duża część ankietowanych zaznaczyła również odpowiedź – nie wiem (36%), a znacznie mniej zgodziło się ze zdaniem, że może to nastąpić do 2027 r. (4,5%) lub po tym roku (10,8%). Należy podkreślić, że są to opinie wyrażone przez osoby, które w dużej części nie zetknęły się z problematyką indeksowania automatycznego.

Ostatnie pytanie, które zostanie uwzględnione w tym porównaniu dotyczyło oceny jakości automatycznego indeksowania. Większość bibliotekarzy w badaniu przeprowadzonym przez szwajcarską badaczkę stwierdziło, że będzie ono gorsze niż opracowanie manualne (N – 59,6%; A – 60,7%). Znacznie mniej ankietowanych zgodziło się, że będzie ono na tym samym poziomie (N – 19,3%; A – 19,7%) lub nie potrafiło udzielić zdecydowanej odpowiedzi (N – 12,3%; A – 16,4%)[66]. Większość polskich bibliotekarzy stwierdziła, że nie potrafi określić jakości opracowania dokumentów przez systemy komputerowe (53,2%). Według 39,6% ankietowanych narzędzia do automatycznego opracowania będą uzyskiwały słabsze rezultaty niż ludzie.

[64] Przykładem projektu wykorzystującego automatyczną ekstrakcję słów kluczowych są prace prowadzone na Politechnice Warszawskiej. Zob.: J. Koperwas [et al.], *Intelligent Information Processing for Building University Knowledge Base*, „Journal of Intelligent Information Systems" 2017, vol. 48, issue 1, s. 141–163.

[65] A. Keller, *op. cit.*, s. 899–901.

[66] *Ibidem*, s. 900–901.

Oceniając uzyskane wyniki należy pamiętać, że indeksowanie automatyczne jest tylko jednym ze sposobów usprawnienia katalogowania. Polskie biblioteki naukowe powszechnie uczestniczą w projekcie współkatalogowania w ramach Katalogu Centralnego NUKAT. Szansą jest również współpraca z wydawcami. Keller wspomina te możliwości i bada opinie na ich temat[67]. Kwestia ta wykracza jednak poza ramy tego artykułu. Sytuacja polskich bibliotekarzy, jeśli chodzi o znajomość problematyki indeksowania automatycznego, jest w dużym stopniu zbliżona do sytuacji w krajach anglojęzycznych. Różnica między tymi dwoma grupami dotyczyła między innymi oceny możliwości zastąpienia indeksowania manualnego przez automatyczne. Polscy bibliotekarze są bardziej konserwatywni w porównaniu z bibliotekarzami anglojęzycznymi. Jednocześnie dużo mniejsza liczba osób stwierdziła, że taki sposób indeksowania będzie charakteryzował się uzyskiwaniem słabszych wyników.

Zakończenie

Osoby wykonujące różne zawody powinny przygotować się na nadchodzące zmiany związane z komputeryzacją. Frey i Osborne pokazali w swoim badaniu, że zmiany te w odmiennym stopniu mogą dotknąć różnych profesji. Przemiany mogą dotyczyć również tych zawodów, które ze względu na brak powtarzalności zadań i liczbę zmiennych wymagających uwzględnienia, do niedawna były zarezerwowane wyłącznie dla ludzi. Bibliotekarze są w grupie tych zawodów, w przypadku których prawdopodobieństwo pełnej komputeryzacji nie jest bardzo wysokie, jednak pewne czynności wykonywane tradycyjnie w sposób manualny mogą zostać w pełni zautomatyzowane. Jedną z nich jest opracowanie zbiorów. W artykule pokazano, że im większe podobieństwo strukturalne dokumentów i węższa dziedzina tym łatwiej będzie skomputeryzować indeksowanie. W tej chwili bardziej prawdopodobne wydaje się wykorzystanie automatycznych narzędzi do opracowania formalnego zbiorów niż do opracowania rzeczowego. Obecnie rozwijane projekty wykorzystują komputery w celu znacznego przyśpieszenia i zwiększenia liczby opracowywanych zasobów.

W kontekście opisywanych przemian istotne są następujące pytania: czy edukacja będzie za nimi nadążać oraz czy pracownicy będą gotowi do zmiany kwalifikacji. Wydaje się, że wskazana byłaby modyfikacja programów nauczania przyszłych bibliotekarzy i specjalistów informacji naukowej, która pozwoliłaby im zdobyć umiejętności umożliwiające wykonywanie i zachowanie

[67] *Ibidem*, s. 901–902.

pracy na coraz bardziej skomputeryzowanym rynku. Należy w nich uwzględnić problematykę automatycznego opracowania zarówno formalnego, jak i rzeczowego tak, aby absolwenci byli przygotowani na zmiany, które nastąpią w przyszłości. Próby wprowadzenia automatycznego czy też raczej semi-automatycznego indeksowania nie oznaczają na razie, że można zrezygnować z edukowania przyszłych specjalistów w zakresie opracowania manualnego zbiorów. Wydaje się, że w najbliższych latach ten sposób opracowania nadal będzie stanowił istotny element pracy w bibliotece. Znaczną przeszkodą w automatyzacji indeksowania jest brak dostępu do wielu pełnych cyfrowych tekstów dokumentów, które nadal muszą być indeksowane przez ludzi. Wspomniane umiejętności mogą być przydatne w procesie planowania projektów oraz sprawdzania wyników pracy komputera. Zalety ich posiadania są widoczne między innymi w przypadku zaprezentowanego projektu MTI.

Oprócz edukowania przyszłych specjalistów informacji istotne jest także kształcenie osób pracujących już w bibliotekach w zakresie korzyści, zalet oraz wad automatycznego indeksowania zbiorów. Przedstawione w artykule badanie ankietowe przeprowadzone wśród bibliotekarzy z polskich bibliotek akademickich pokazuje, że pracownicy tych instytucji mają niewielką wiedzę w tym zakresie. Należy przy tym podkreślić, że ich opinie nie odbiegają znacznie od postaw bibliotekarzy anglojęzycznych. W podsumowaniu swojego badania Keller wskazała na konieczność współpracy bibliotekarzy z informatykami, która powinna być oparta na próbie wzajemnego zrozumienia i chęci dzielenia się wiedzą[68]. Bibliotekarze powinni mieć zapewniony udział w projektach z zakresu automatycznego indeksowania jako specjaliści od jakości metadanych i od organizacji zbiorów. Jednocześnie należy pamiętać o dynamicznym rozwoju wyszukiwania pełnotekstowego coraz częściej wspomaganego przez bazy wiedzy[69].

W artykule podkreślono ograniczony charakter przeprowadzonej ankiety. Było to badanie wstępne, które umożliwiło jedynie zarysowanie ogólnego obrazu bez uwzględnienia wielu istotnych szczegółów. Środowisko bibliotekarskie od wielu lat wykorzystuje w swojej pracy nowoczesne narzędzia. Ważne pytanie dotyczy zakresu znajomości nowych technologii wśród bibliotekarzy. Większe badanie mogłoby uwzględniać nie tylko automatyczne indeksowanie, ale również na przykład problematykę technologii mobilnych czy też

[68] *Ibidem*, s. 904.

[69] W artykule wspomniano o wyszukiwarce naukowej Microsoft Academic, która wykorzystuje Microsoft Academic Graph. Bazy wiedzy są budowane również przez inne przedsiębiorstwa z branży wyszukiwarek internetowych. Zob.: A. Singhal, *Introducing the Knowledge Graph: Things, Not Strings*, https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html, [dostęp: 6.04.2017].

chmur obliczeniowych. Należałoby włączyć do badania bibliotekarzy z różnych typów bibliotek, aby uzyskać szerszy obraz wiedzy i potrzeb środowiska. Wyniki mogłyby być punktem wyjścia do modyfikacji programów nauczania tak, aby dostosować je do zmieniających się warunków pracy bibliotekarskiej w obliczu ciągle postępującej komputeryzacji. Istotnym elementem tych przemian jest automatyczne lub semi-automatyczne indeksowanie zbiorów, na które należy uwagę wobec widocznego postępu algorytmów indeksujących przynoszących coraz lepsze efekty.

Podziękowania

Chciałbym podziękować wszystkim bibliotekarzom z bibliotek akademickich, którzy poświęcili swój czas na wypełnienie ankiety wykorzystanej w artykule.

Bibliografia

De Domenico M., Omodei E., Arenas A., *Quantifying the Diaspora of Knowledge in the Last Century*, http://arxiv.org/abs/1604.00696, [dostęp: 6.04.2017].

Demner-Fushman D., Mork J., *NLM Medical Text Indexer Technical Report to the LHNCBC Board of Scientific Counselors April 2016*, Bethesda 2016, https://lhncbc.nlm.nih.gov/system/files/pub9359.pdf, [dostęp: 6.04.2017].

Fact Sheet MEDLINE, PubMed and PMC (PubMed Central): How are they different, https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html, [dostęp: 6.04.2017].

Formularze Google, [dostęp: 6.04.2017], https://www.google.pl/intl/pl/forms/about/.

Frey C.B., Osborne M.A., *The Future of Employment: How Susceptible are Jobs to Computerisation?*, „Technological Forecasting and Social Change" 2017, vol. 114, s. 254–280.

Glasser S., *The Changing Face of Cataloging Positions at Academic Institutions: What Skill Set is Needed, and How Can Students Prepare?*, „The Serials Librarian" 2007, vol. 51, issue 3–4, s. 39–49.

Golub K. [et al.], *A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval*, „Journal of the Association for Information Science and Technology" 2016, vol. 67, issue 1, s. 3–16.

Gregor D., Mandel C., *Cataloging Must Change!*, „Library Journal" 1991, vol. 116, issue 6, s. 42–47.

Hug S.E., Ochsner M., Brändle M.P., *Citation Analysis with Microsoft Academic*, „Scientometrics" 2017, vol. 111, issue 1, s. 371–378.

Inclusion Guidelines for Webmasters: Indexing Guidelines, https://scholar.google.com/intl/en/scholar/inclusion.html#indexing, [dostęp: 6.04.2017].

Indeksowanie treści w teorii i praktyce: warsztaty, http://www.novaskills.pl/docs/Indeksowanie_tresci.pdf, [dostęp: 6.04.2017].

Junger U., *Can Indexing Be Automated? The Example of the Deutsche Nationalbibliothek*, „Cataloging & Classification Quarterly" 2014, vol. 52, issue 1, s. 102–109.

Keller A., *Attitudes among German- and English-Speaking Librarians toward (Automatic) Subject Indexing*, „Cataloging & Classification Quarterly" 2015, vol. 53, issue 8, s. 895–904.

Key MEDLINE Indicators, https://www.nlm.nih.gov/bsd/bsd_key.html, [dostęp: 6.04. 2017].

Keyser P. de, *Indexingn From Thesauri to the Semantic Web*, Oxford 2012.

Kont K.-R., *How Much Does It Cost to Catalog a Document? A Case Study in Estonian University Libraries*, „Cataloging & Classification Quarterly" 2015, vol. 53, issue 7, s. 825–850.

Koperwas, J. [et al.], *Intelligent Information Processing for Building University Knowledge Base*, „Journal of Intelligent Information Systems" 2017, vol. 48, issue 1, s. 141–163.

Malak P., *Indeksowanie treści: porównanie skuteczności metod tradycyjnych i automatycznych*, Warszawa 2012.

Mann T., *"Cataloging Must Change!" and Indexer Consistency Studies: Misreading the Evidence at Our Peril*, „Cataloging & Classification Quarterly" 1997, vol. 23, issue 3–4, s. 3–45.

Mapy tematów jako system reprezentacji wiedzy, http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.cejsh-b170164a-93d9-4324-80d9-d99351d718dc, [dostęp: 6.04.2017].

Mendeley | How Does the Automatic Document Details Extraction Work, http://support.mendeley.com/customer/en/portal/articles/227883-how-does-the-automatic-document-details-extraction-work-, [dostęp: 6.04.2017].

MeSH on Demand, https://www.nlm.nih.gov/mesh/MeSHonDemand.html, [dostęp: 6.04.2017].

Mork J., Aronson A., Demner-Fushman D., *12 Years on – Is the NLM Medical Text Indexer Still Useful and Relevant?*, „Journal of Biomedical Semantics" 2017, vol. 8, s. 1–10.

MTI First Line (MTIFL) Indexing, https://ii.nlm.nih.gov/MTI/MTIFL_Journal_List.pdf, [dostęp: 6.04.2017].

Ortega J.L., *Academic Search Engines: A Quantitative Outlook*, Amsterdam 2014.

Park J., Brenza A., *Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art*, „Information Technology & Libraries" 2015, vol. 34, issue 3, s. 22–42.

Park J., Tosaka Y., *Advancing Professional Learning in Libraries: An Exploratory Study of Cataloging and Metadata Professionals' Experiences and Perspectives on Continuing Education Issues*, „Cataloging & Classification Quarterly" 2017, vol. 55, issue 3, s. 153–171.

Pofeldt E., *Study: Will A Robot Do Your Job Some Day?*, https://www.forbes.com/sites/elainepofeldt/2014/02/26/will-r2-d2-snag-your-job/#716f4ad51897, [dostęp: 6.04.2017].

Pulgarín A., Gil-Leiva I., *Bibliometric Analysis of the Automatic Indexing Literature: 1956–2000*, „Information Processing & Management" 2004, vol. 40, issue 2, s. 365–377.

Randtke W., *Automated Metadata Creation: Possibilities and Pitfalls*, „The Serials Librarian" 2013, vol. 64, issue 1–4, s. 267–284.

Singhal A., *Introducing the Knowledge Graph: Things, Not Strings*, https://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-thingsnot.html, [dostęp: 6.04.2017].

Sinha A. [et al.], *An Overview of Microsoft Academic Service (MAS) and Applications*, [w:] *Proceedings of the 24th International Conference on World Wide Web Companion (WWW 2015 Companion)*, New York 2015, s. 243–246, https://doi.org/10.1145/2740908.2742839, [dostęp: 6.04.2017].

Stylianou N. [et al.], *Will a Robot Take Your Job?*, http://www.bbc.com/news/technology-34066941, [dostęp: 6.04.2017].

Wu J. [et al.], *CiteSeerX Data: Semanticizing Scholarly Papers*, [w:] *Proceedings of the International Workshop on Semantic Big Data*, New York 2016, s. 1–6, https://doi.org/10.1145/2928294.2928306, [dostęp: 6.04.2017].

Zavorin I., Mork J.G., Demner-Fushman D., *Using Learning-To-Rank to Enhance NLM Medical Text Indexer Results*, [w:] *Proceedings of the Fourth BioASQ Workshop*, s. 8–15, http://aclweb.org/anthology/W/W16/W16-3102.pdf, [dostęp: 06.04.2017].