DOI 10.4467/0023589XKHNT.25.001.21319

Maria Guzik-Jureczka
Instytut Historii im. T. Manteuffla PAN
ORCID 0009-0008-1722-2001

Piotr Jaskulski
Instytut Historii im. T. Manteuffla PAN
ORCID 0000-0002-3616-4643

Adam Zapała
Instytut Historii im. T. Manteuffla PAN
ORCID 0000-0003-3450-8286

Rola narzędzi cyfrowych w przetwarzaniu i udostępnianiu biogramów postaci historycznych

The Role of Digital Tools in Processing and Sharing Biographical Entries of Historical Figures

The dynamic growth of biographical research in recent years, both in traditional and digital forms, has led to an immense increase in available data but has also created the problem of repeated information about the same figures appearing in various works. This often forces researchers to search through dozens of biographical entries and databases to obtain a complete picture of a person's life, which is time-consuming and creates interpretative challenges. This article proposes the use of digital methods, including the FAIR principles, Linked Open Data (LOD) standards, and Large Language Models (LLMs), to facilitate the access, exchange, and reuse of biographical data. It discusses methods that allow for the harmonization and integration of biographical data with reference databases, as well as solutions that enable browsing and sharing information from large text corpora using artificial intelligence and chatbot interfaces. The study also highlights the changes introduced by digital research methods, such as the separation of content from form and the possibility of diverse visualization of data, along with the need for researchers to change their way of thinking to align with the principles of open science.

Keywords: biographical research, digital humanities, artificial intelligence (Large Language Models), FAIR, Linked Open Data

Słowa kluczowe: biografistyka, humanistyka cyfrowa, sztuczna inteligencja (duże modele językowe), FAIR, Linked Open Data

Ostatnie lata przyniosły bardzo dynamiczny rozwój przedsięwzięć w dziedzinie biografistyki, powodując ogromny przyrost materiału. Chodzi zarówno o nowe przedsięwzięcia słownikowe[1], monografie zawierające biogramy[2], jak i bazy danych[3]. W przypadku tych ostatnich mamy do czynienia zarówno z przedsięwzięciami w pełni naukowymi, jak i o charakterze hobbystycznym i popularyzatorskim[4]. Różne zasoby często powtarzają informacje (np. biogramy tych samych osób w Polskim Słowniku Biograficznym (PSB) i innych słownikach). Uzyskanie pełnej wiedzy na temat życiorysu jakiejś postaci wymaga często przeszukania i wykorzystania dziesiątek biogramów i baz danych. Z perspektywy użytkownika sprawia to wrażenie intelektualnej inflacji, gdzie wysiłek konieczny do przeszukania całości zasobów biograficznych jest nieproporcjonalnie większy niż wartość informacji pozyskanych w wyniku takiej kwerendy. Co gorsza, taka sytuacja sprawia, że często wyszukane informacje są niepewne lub zgoła wykluczające się (np. wiele dat ze starszych tomów PSB zdezaktualizowała się dzięki nowym badaniom, przez co informacje odnalezione w trakcie kwerendy mogą się różnić w zależności od źródła). W takiej sytuacji czytelnik musi często sam zweryfikować pozyskaną wiedzę i zdecydować, któremu opracowaniu wierzyć. Celem niniejszego tekstu jest opisanie możliwości wykorzystania metod cyfrowych do rozwiązania tego typu problemów.

Rozważania niniejsze należy rozpocząć od nakreślenia fundamentalnych zmian, jakie metody cyfrowe oraz powszechny dostęp do Internetu wprowadziły w dziedzinie upowszechniania i wymiany wiedzy. Pie rwszą diametralną zmianą jest oddzielnie formy przekazu od jej treści. O ile w przypadku tradycyjnego pisarstwa historycznego treść i forma były ze sobą nierozerwalnie związane, o tyle w przypadku informacji modelowanych w środowisku cyfrowym oba te zagadnienia są całkowicie oddzielne. Treścią są informacje przechowywane w bazie danych, natomiast formą jest sposób ich prezentacji, zapewniany przez dostosowane do potrzeb interfejsy i aplikacje. Jeżeli w przypadku tradycyjnych tekstów informacje w nich zawarte mogły być wydobyte tylko przez ich odczytanie, to w przypadku projektów cyfrowych nie jest żadnym problemem (a jest wręcz normą), by te same dane prezentować w różny sposób, np. w formie tekstu, mapy, diagramu, tabeli. Co więcej, te same dane mogą być wykorzystane w kontekście całkowicie odmiennym od przewidzianego podczas ich tworzenia.

To rozróżnienie prowadzi do diametralnej zmiany w kwestii ponownego wykorzystania informacji. W przypadku tekstów drukowanych kopiowanie, ponowne wykorzystywanie i upowszechnianie tego samego tekstu jest plagiatem, a więc działaniem nieetycznym oraz karalnym. Nieco inaczej sprawa ta wygląda w przypadku projektów cyfrowych. Oczywiście nieautoryzowane wykorzystanie czyjejś pracy (np. danych) bez wskazania źródła pozostaje plagiatem, jednakże charakter projektów cyfrowych sprawia, że twórcy baz da-

1 Poza wydawanym od 90 lat Polskim Słownikiem Biograficznym badacze dysponują dzisiaj całym wachlarzem słowników specjalistycznych, takich jak np.: *Słownik badaczy literatury polskiej*, t. 1–10, red. J. Starnawski, Łódź 1994–2009; *Słownik biograficzny historyków łódzkich*, red. J. Kita, R. Stobiecki, Łódź 2000; *Słownik techników polskich*, t. 1–30, Warszawa 1989–2024 itd.

2 Do tej grupy należą m.in. opracowania prozopograficzne takie jak np.: P. Dembiński, *Poznańska kapituła katedralna schyłku wieków średnich*, Poznań 2012.

3 Przykładem biograficznej bazy danych jest *Corpus Academicum Cracoviense*, cac.historia.uj.edu.pl [dostęp 20.09.2024].

4 Oprócz przedsięwzięć takich jak Wikipedia czy Wikidata należy wymienić np. portale genealogiczne, takie jak np.: Maria Jadwiga Minakowska, *Genealogia Potomków Sejmu Wielkiego*, www.sejm-wielki.pl [dostęp 20.04.2024].

nych często nie mają nic przeciwko ponownemu wykorzystaniu ich zasobów lub wręcz dążą do niego. Z tego też powodu większość projektów cyfrowych udostępniana jest nieodpłatnie na zasadzie otwartego dostępu na wybranej licencji *open source*[5]. W zależności od intencji autor może zdecydować się na przyznanie użytkownikowi różnych praw. Istnieją licencje pozwalające jedynie na kopiowanie i upowszechnianie[6], ale także bardziej liberalne dające użytkownikom prawo dowolnego zmieniania treści dzieła jedynie z koniecznością udostępnienia zmienionej wersji na tych samych zasadach ze wskazaniem autora oryginalnego zasobu[7]. Otwartość Internetu sprawia, że nierzadko autorom bardziej zależy na wskazaniu ich jako twórców niż na zamknięciu dostępu do danego zasobu. Takie podejście całkowicie zmienia podstawy dyskursu naukowego. Z jednej strony wymusza dowartościowanie zasobów cyfrowych i podawanie odwołań do nich (co często tradycyjnym historykom wydaje się zbędne), z drugiej natomiast pozwala na agregację i ponowne wykorzystywanie danych.

Druga niezmiernie ważna zmiana wynika ze skutecznego działania chatbotów wykorzystujących duże modele językowe (najbardziej znany to ChatGPT), pozwalających na automatyczne generowanie tekstu[8]. Teksty wygenerowane przez GPT niewiele różnią się od napisanych przez człowieka, co owocuje coraz szerszym wykorzystaniem maszyny do pisania tekstów. Wysoka jakość rezultatów sprawia coraz większe problemy w odróżnieniu tekstu wygenerowanego automatycznie od napisanego przez człowieka, a co za tym idzie zdefiniowaniu i wskazaniu autorstwa[9]. Za tą zmianą musi podążać dewaluacja generowania tekstu jako czynności twórczej, skoro może ją bezbłędnie robić komputer. Całościowe konsekwencje tej zmiany nie są jeszcze znane, ale można przypuszczać, że będzie ona miała rewolucyjne znaczenie dla ludzkiej kultury. Skoro komputer potrafi bezbłędnie generować tekst w formie trudnej do odróżnienia od twórczości człowieka, to uzasadnione staje się pytanie o miejsce pisarstwa (w tym pisarstwa historycznego) w warsztacie historyka.

Obie wyżej wymienione kwestie wymuszają podjęcie bardzo głębokiej refleksji i reformy myślenia o fundamentach pracy badacza historii (w tym także zajmującego się biografistyką). Wdrożenie cyfrowych metod badawczych w dziedzinie historii zwiastuje nową erę w modelowaniu i dystrybucji informacji, przewyższając wydajnością tradycyjne podejścia. Ten postęp, choć napotyka na pewien opór ze strony zwolenników klasycznych metod, wydaje się w sposób nieunikniony zmierzać ku triumfowi nowych technologii, oferując obiecujące perspektywy dla przyszłości badań historycznych. Pojawia się

5 Udostępnianie zasobów w Internecie często odbywa się na licencjach Creative Commons, creativecommons.org [dostęp 8.09.2024]

6 Licencja CC BY-ND-NC 4.0: *Attribution-NonCommercial-NoDerivatives 4.0 International*, creativecommons.org/licenses/by-nc-nd/4.0 [dostęp 20.09.2024].

7 Licencja CC BY-NC 4.0: *Attribution-NonCommercial 4.0 International*, creativecommons.org/licenses/by-nc/4.0 [dostęp 13.09.2024], lub w przypadku oprogramowania GNU General Public Licence, www.gnu.org/licenses/gpl-3.0.html [dostęp 8.09.2024].

8 Systemy dialogowe oparte na AI stały się obecnie bardzo popularne. Oprócz ChatGPT firmy OpenAI swoje rozwiązanie oferuje również Google w postaci chata Google Gemini. Uznanie użytkowników zdobywa też system Anthropic Claude. Wiele z nich oprócz dialogu z użytkownikiem jest w stanie także przeczytać i przeanalizować dostarczone pliki czy przeszukiwać Internet w celu udzielenia lepszych odpowiedzi. Oprócz rozwiązań komercyjnych rozwijane są także duże modele językowe *open source*. Mniejsze z nich jak np. Llama-3.1-8B czy polski Bielik-2 można uruchomić lokalnie nawet na laptopie, korzystając z narzędzi typu LM Studio.

9 Powstają liczne systemy wykrywające użycie AI w tekście, jednak ich skuteczność nie jest obecnie idealna.

pytanie, jak ten model upowszechniania informacji zaadaptować z pożytkiem dla nauki i badaczy.

FAIR i Linked Open Data

W przypadku biografistyki, jako gałęzi badań historycznych szczególnie ważnej dla społeczeństwa, rzeczą niepodlegającą dyskusji jest konieczność zastosowania nowoczesnych standardów wymiany wiedzy. Zmiana musi zajść zarówno na poziomie modelowania treści materiałów naukowych, jak i ich formy udostępnienia. Kluczowymi zagadnieniami z perspektywy modelowania danych są: pryncypia FAIR[10] oraz standardy Linked Open Data (LOD)[11].

FAIR (*Findable, Accessible, Interoperable, Reusable*) są to zasady, według których należy udostępniać przygotowywane dane i informacje, by umożliwić do nich dostęp innym użytkownikom (np. badaczom). W 2016 r. zostały one dokładnie zdefiniowane i opisane wraz z wytycznymi, jak rozumieć każdy z komponentów[12]. *Findable* oznacza „możliwy do znalezienia"; dotyczy to zarówno danych, jak i opisujących je metadanych. Te informacje powinny posiadać trwały, globalnie unikalny identyfikator oraz być udostępnione w bazie, która pozwala na przeszukiwanie zgromadzonych zasobów. Kolejną zasadą jest *Accessible*, czyli „dostępny". W praktyce oznacza to użycie standardowego protokołu komunikacyjnego do udostępnienia danych. Protokół ten musi być bezpłatny i otwarty, a także w razie konieczności zapewniać procedurę uwierzytelnienia i autoryzacji dla jego użytkowników. Ta zasada wymaga również, aby metadane zasobu pozostały dostępne, nawet jeśli sam zasób przestanie być osiągalny. Poza tym dane powinny być „interoperacyjne" – *Interoperable* – a więc możliwe do połączenia z innymi danymi i zawierające odniesienia do zewnętrznych źródeł wiedzy. Do opisu informacji powinny zostać wykorzystane formalne i ogólnie przyjęte standardy reprezentacji wiedzy oraz słowniki, które same są zgodne z pryncypiami FAIR. Ostatnią z zasad jest „możliwy do ponownego wykorzystania" – *Reusable*. Do spełnienia tego wymogu potrzebne są informacje o autorze oraz wyczerpujący opis zasobu pozwalający dobrze zrozumieć jego zawartość. Dane powinny być również udostępnione z odpowiednią licencją określającą warunki ponownego użycia.

Przygotowanie danych zgodnie z zasadami FAIR niesie za sobą wiele korzyści. Informacje zgromadzone w ten sposób są nie tylko czytelne dla człowieka, ale też możliwe do przetwarzania przez programy komputerowe. Pozwala to np. na automatyczne przeszukiwanie dużych zbiorów na podstawie informacji zawartych w ich metadanych. Użycie standardów sprawia, że informacje są dobrze dostępne i łatwe do zrozumienia przez kolejnych użytkowników, a więc w wygodny sposób mogą się przysłużyć kolejnym badaniom bez konieczności powielania już wykonanej pracy.

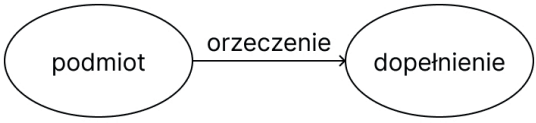
Dane przygotowane zgodnie z pryncypiami FAIR można następnie powiązać w ramach standardu LOD. Określa on zasady publikacji danych w Internecie tak, aby były one do-

10 *FAIR Principles*, www.go-fair.org/fair-principles/ [dostęp 30.01.2024].
11 *LinkedData*, www.w3.org/wiki/LinkedData [dostęp 30.01.2024].
12 M.D. Wilkinson, M. Dumontier, I. Aalbersberg i in., *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, „Scientific Data" 2016, t. 3, 160018.

stępne w sposób otwarty oraz wzajemnie ze sobą połączony. Udostępnianie materiałów badawczych za pomocą zdefiniowanych przez LOD struktur pozwala na łatwe powiązanie ich z wcześniej opublikowanymi danymi oraz informacjami z innych źródeł. Sprzyja to popularyzacji wiedzy naukowej oraz znacznie ułatwia przeprowadzanie badań dzięki możliwości wyszukiwania połączeń między danymi. Informacje dostępne w alternatywnych źródłach mogą też pozwolić na automatyczne wyciąganie wniosków o powiązanych z nimi danych.

W celu wzajemnego powiązania bytów (np. osób, miejsc, ale też całych zasobów) konieczne jest, aby każdy z nich można było jednoznacznie zidentyfikować. Do takiej identyfikacji służą Ujednolicone Identyfikatory Zasobów (*Uniform Resource Identifier*, URI). Przykładem URI jest URL (*Uniform Resource Locator*), czyli dobrze znany wszystkim adres strony internetowej. Szczególną wartość mają Trwałe Identyfikatory (*Persistent Identifier*, PI), które po przydzieleniu nigdy nie zostaną zmienione i muszą być globalnie unikalne[13] – nie tylko w obrębie danej bazy, lecz całego Internetu. Możliwość nadawania Trwałych Identyfikatorów zapewnia np. system Handle (wykorzystywany m.in. do nadawania identyfikatorów DOI)[14].

LOD określa pięć kolejnych stopni oceny jakości udostępnianych danych. Aby uzyskać najwyższą ocenę, dane muszą być dostępne w otwartym, ustrukturyzowanym i niezastrzeżonym formacie, opatrzone URI, połączone z innymi otwartymi źródłami oraz zgodne ze standardami rekomendowanymi przez World Wide Web Consortium (W3C)[15]. W tej dziedzinie jedną z najistotniejszych wytycznych jest stosowanie modelu RDF[16] (*Resource Description Framework*), czyli sposobu wyrażania informacji w postaci trójek: podmiot – orzeczenie – dopełnienie (*subject – predicate – object*). Podmiotem jest określany byt, orzeczenie (predykat) wskazuje nam na pewną jego własność, a dopełnienie (obiekt) jest wartością tej własności. Rozszerzając informacje o zasobie o kolejne cechy, nadajemy im strukturę grafu skierowanego, którego węzłami są odrębne byty, a krawędziami relacje pomiędzy nimi. Dane przechowywane w tak ściśle ustrukturyzowany sposób mogą być następnie przeszukiwane z użyciem zapytań w języku SPARQL. Język ten pozwala również na manipulację danymi.



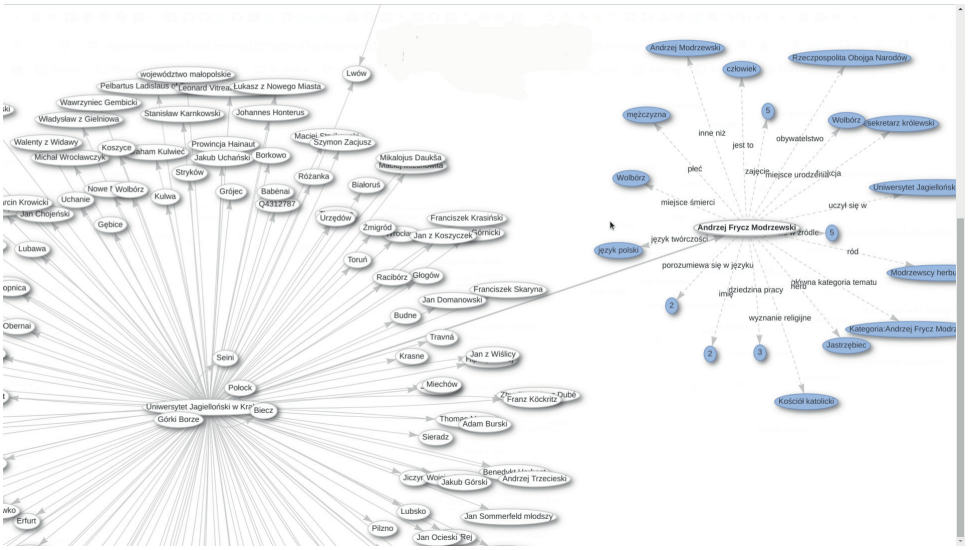
Ryc. 1. Model RDF: trójka podmiot – orzeczenie – dopełnienie (rysunek autorów)

13 K. Richards, R. White, N. Nicolson, R. Pyle, *A Beginner's Guide to Persistent Identifiers*, www.gbif.org/document/80575/a-beginners-guide-to-persistent-identifiers [dostęp 24.09.2024].

14 Jest to system służący do przydzielania identyfikatorów obiektom cyfrowym zarządzany przez Corporation for National Research Initiatives (CNRI). Jedną z implementacji systemu Handle jest DOI (*Digital Object Identifier*), wykorzystywany do identyfikacji m.in. publikacji naukowych. Więcej: *HDL.NET® Information Services*, handle.net [dostęp 11.12.2023].

15 *W3C*, www.w3.org [dostęp 11.12.2023]

16 *Resource Description Framework (RDF)*, www.w3.org/RDF [dostęp 11.12.2023]



Ryc. 2. Przykład wizualizacji danych biograficznych w formacie RDF (źródło: Wikidata.org)

Zasoby spełniające zasady określone w standardzie LOD tworzą wspólnie Sieć Semantyczną (*Semantic Web*). Wizualnie można ją przedstawić jako graf zbudowany z poszczególnych zbiorów danych, takich jak np. bazy powstałe z wykorzystaniem oprogramowania Wikibase. Dzięki połączeniom między zasobami Sieć Semantyczna umożliwia swobodne przechodzenie pomiędzy danymi udostępnionymi przez niezależne jednostki i organizacje. Daje to ogromne możliwości jednoczesnego wydobywania informacji z różnych źródeł, bez konieczności powielania ich w każdym ze zbiorów. Jest to podejście zupełnie różne od sposobu przechowywania i udostępniania informacji za pomocą dokumentów pisanych językiem naturalnym. Tekst jest co prawda zrozumiały i czytelny dla odbiorców, ale cechuje go duża objętość w stosunku do ilości informacji w nim zawartych oraz częsty brak połączeń pomiędzy różnymi źródłami danych dotyczących tego samego tematu. Powoduje to trudności w trakcie wyszukiwania konkretnych zasobów i sprawia, że ten proces jest bardzo czasochłonny. Sieć Semantyczna rozwiązuje te problemy i gwarantuje, że udostępnione w niej dane mogą być przetwarzane maszynowo przy jednoczesnym zachowaniu czytelności dla ludzi.

Wykorzystanie Sieci Semantycznej nie byłoby możliwe bez systemu jednoznacznego definiowania. Tylko działając na danych umiejscowionych w ściśle określonym i zhierarchizowanym modelu, możemy zbiorczo przetwarzać zgromadzoną wiedzę. Odpowiednią strukturę dla zasobów możemy zapewnić wykorzystując ontologie dziedzinowe[17], SKOS (*Simple Knowledge Organization System*)[18] lub słownictwo kontrolowane. Wszystkie te rozwiązania wyznaczają zakres informacji, który jest możliwy do wyrażenia w danym systemie. Wśród nich ontologie dziedzinowe dostarczają najpełniejszy i najbardziej formalny

17 Ontologie wyrażone są w języku OWL: *Web Ontology Language (OWL)*, www.w3.org/OWL [dostęp 30.01.2024].

18 A. Miles, B. Matthews, D. Beckett, D. Brickley, M. Wilson, N. Rogers, *SKOS: A Language to Describe Simple Knowledge Structures for the Web*, epubs.stfc.ac.uk/manifestation/685/SKOS-XTech2005.pdf [dostęp 20.09.2024].

opis wiedzy składający się ze zbioru pojęć oraz łączących je relacji, wyrażonych za pomocą logiki formalnej. Nieco mniej rozbudowane schematy można tworzyć wykorzystując SKOS, umożliwiający reprezentowanie tezaurusów, taksonomii oraz innych podobnych struktur służących do organizacji informacji. Tego rodzaju systemy obejmują zwykle węższy zakres niż ontologie i nie zapewniają tak samo wysokiego poziomu semantyki formalnej, a co za tym idzie równie zaawansowanych możliwości logicznego wnioskowania. Najprostszym narzędziem jest słownictwo kontrolowane, czyli ściśle sprecyzowany zbiór możliwych do wykorzystania słów. Wybór jednego z tych rozwiązań zależy od pożądanego stopnia formalności systemu oraz zakresu prezentowanych danych. Każde z nich pozwoli jednak na przekazanie wiedzy w sposób możliwy do zrozumienia i przetworzenia przez programy komputerowe.

Wykorzystanie przedstawionych standardów i mechanizmów przy udostępnianiu danych w Sieci Semantycznej skutkuje ich wprowadzeniem do globalnego obiegu. Jest to niewątpliwie ogromną zaletą takiego sposobu upubliczniania informacji, nieosiągalną dla tradycyjnych źródeł pisanych. Przyporządkowując udostępniane zasoby do bytów zdefiniowanych w istniejących modelach i ontologiach dziedzinowych ułatwiamy dostęp do nich badaczom z całego świata. Po wprowadzeniu danych do wybranej bazy będącej węzłem Linked Open Data możemy powiązać je poprzez identyfikatory np. z danymi zagranicznymi (takimi jak VIAF[19]), a następnie korzystać z informacji obecnych w tych źródłach bez konieczności kopiowania ich i wprowadzania do własnych zasobów.

W jaki sposób zaaplikować wyżej opisane rozwiązania w badaniach biograficznych? Pierwszym krokiem musi być przejście projektów historycznych na poziom danych. Kurczowe trzymanie się druku jako głównego sposobu upowszechniania informacji jest dziś przeciwskuteczne. Można je porównać z jazdą konną. Jest ona być może elegancka, jednak chcąc szybko dotrzeć w jakieś miejsce, wybierzemy raczej bardziej nowoczesny środek transportu. Tak samo musi się stać z nauką. Aby dostarczać dane wysokiej jakości, trzeba uznać, że to przygotowanie informacji i danych jest głównym celem historycznych projektów badawczych. Ich analiza i interpretacja pozwala wydobywać wiedzę i osiągnąć mądrość[20].

Jeśli zaakceptujemy konieczność wytwarzania danych, należy przedyskutować sposób ich modelowania. W tym przypadku bezsprzecznie konieczne jest trzymanie się pryncypiów FAIR i standardów LOD. Dane więc powinny być opisane metadanymi, zapisane w formacie pozwalającym na ich ponowne przetworzenie, zamodelowane przy użyciu standardów (ontologii, słownictwa kontrolowanego etc.). Udostępnienie danych w taki sposób jest dużo łatwiejsze do opisania niż do zrealizowania. Potrzebna do tego jest infrastruktura sprzętowa (serwery), programistyczna (środowisko), a także wiedza. Zapewnienie wszystkich tych trzech aspektów jest celem infrastruktury Dariah.lab[21]. W jej ra-

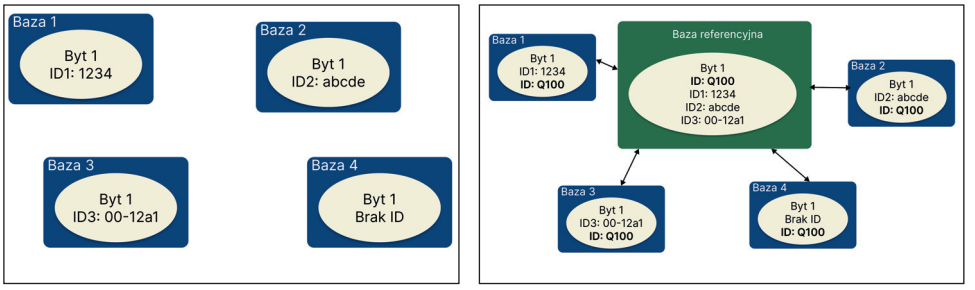
19 VIAF. Virtual International Authority File, viaf.org [dostęp 11.12.2023].

20 Wiedza i Mądrość to ważne kategorie w inżynierii informacji. Obszerne definicje tych pojęć znaleźć można w książce Bogdana Stefanowicza, *Informacja. Wiedza. Mądrość*, Warszawa 2013 (Biblioteka Wiadomości Statystycznych t. 66).

21 Infrastruktura została zbudowana przez konsorcjum 15 czołowych ośrodków akademickich w Polsce zajmujących się humanistyką cyfrową: *Dariah.lab*, lab.dariah.pl, [dostęp 30.01.2024]. Za przygotowanie rozwiązań dopasowanych do potrzeb badań historycznych odpowiada Instytut Historii Polskiej Akademii Nauk: *Dariah.lab*, ihpan.edu.pl/dariah-lab [dostęp 8.09.2024].

mach utworzono bazę wiedzy dla polskich danych historycznych, WikiHum[22]. Obecnie są w niej zgromadzone głównie informacje o miejscach i osobach ważnych dla historii Polski, może ona jednak zostać rozbudowana o dowolne informacje. Baza ta zapewnia Trwałe Identyfikatory (Identyfikator Dariah.lab), a także przyjazne w użyciu narzędzia do rekoncyliacji, przeszukiwania i wizualizacji danych. Jej model danych oparty jest na standardach (np. ontologii Ontohgis), a także modelu danych Wikidata (najważniejszego węzła LOD). Wykorzystanie Trwałych Identyfikatorów oraz narzędzia do rekoncyliacji[23] pozwala na łatwą harmonizację innych danych z bazą WikiHum.

Nadawanie Trwałych Identyfikatorów danym i łatwość wiązania WikiHum z innymi zasobami daje możliwość wykorzystania jej jako bazy referencyjnej. Baza taka może służyć jako źródło jednoznacznej identyfikacji dla innych zasobów, które dzięki wykorzystaniu jej identyfikatorów byłyby wzajemnie ze sobą powiązane. Baza referencyjna musiałaby też być cały czas rozwijana, tak aby każdy nowy rekord w innej bazie danych miał w niej swój odpowiednik.



a. b.

Ryc. 3. Wiązanie danych z różnych źródeł przy pomocy identyfikatorów bazy referencyjnej: a. relacje bez użycia bazy referencyjnej (a raczej ich brak); b. relacje przy użyciu bazy referencyjnej (autorzy: Adam Zapała, Maria Guzik-Jureczka)

Jaki wpływ opisany wyżej sposób udostępniania informacji ma na pracę z danymi biograficznymi i pisanie haseł biograficznych? Wymusza on poza pisaniem tekstów także przygotowanie zestawów danych, które powinny być harmonizowane, a częściowo też integrowane z bazą referencyjną (np. WikiHum). Harmonizacja polegałaby na uzupełnieniu w lokalnej bazie danych (np. bazie PSB) właściwości „identyfikator Dariah.lab". W ten sposób użytkownik (i co ważniejsze komputer) wiedziałby, że dana osoba z bazy lokalnej to odpowiednia osoba w bazie referencyjnej. Integracja polegałaby na zaciągnięciu do bazy lokalnej brakujących informacji identyfikujących z bazy referencyjnej, a także dodaniu do bazy referencyjnej identyfikatora bazy lokalnej. Co ważne, nie wszystkie dane musiałyby zostać skopiowane, gdyż system komputerowy, wiedząc, że rekordy w różnych bazach dotyczą tej samej postaci, mógłby automatycznie przeszukać wszystkie połączone zasoby. Choć w rozwiązaniu tym powiązania są bilateralne (baza lokalna – baza referencyjna), to w rzeczywistości wykorzystanie identyfikatorów bazy referencyjnej pozwala na wiązanie ze sobą wszystkich baz je wykorzystujących. Zadanie pytania o byt o identyfikatorze Q100 do wszystkich tych baz i każdej z osobna zwróci ten sam byt w odpowiedzi.

22 *WikiHum*, wikihum.lab.dariah.pl/wiki/Main_Page [dostęp 30.01.2024].
23 *OpenRefine*, openrefine.apps.paas.psnc.pl [dostęp 30.01.2024].

Takie rozwiązanie można zaaplikować nie tylko do nowo tworzonych biogramów, lecz także do tekstów istniejących w obiegu w formie tradycyjnej. W tym przypadku możliwe jest przemodelowanie tekstów do form bazodanowych (np. przez anotację i automatyczne wydobywanie bytów z tekstu – NER) lub też przez stworzenie z nich cyfrowego korpusów tekstów (opisanych metadanymi), do których można by jednoznacznie referować (tzn. zastosować bezpośrednie unikalne odniesienia) z wykorzystaniem Trwałych Identyfikatorów. Przemodelowanie zasobów biograficznych do tej formy pozwoliłoby zainteresowanemu badaczowi na łatwe zadanie pytania do bazy, która dałaby mu najważniejsze informacje o obiekcie i listę wszystkich referencji (bazodanowych i tekstowych). Zastosowanie wspólnych standardów (np. ontologii, słowników) oraz otwartych protokołów komunikacji (za pomocą Interfejsu Programowania Aplikacji, API[24]) pozwoliłoby też na tworzenie agregatorów, które dawałyby możliwość przeszukiwania w jednym miejscu metadanych różnych zasobów. Przykładem takiego agregatora jest Federacja Bibliotek Cyfrowych[25]. Pozwala ona przeszukiwać w jednym miejscu metadane wielu połączonych bibliotek cyfrowych i w odpowiedzi na zapytanie udostępniać linki do zgromadzonych w nich zasobów. System działający na podobnych zasadach można zastosować także w przypadku danych biograficznych.

Duże modele językowe

Jeszcze większą zmianę nowe technologie wprowadzają w kwestii formy przygotowywania publikacji biograficznych. W tym przypadku poza tradycyjnym tekstem, można wybrać także inne wizualizacje danych (np. na mapach, diagramach, tabelach), jednak rewolucję w tym zakresie wprowadza wykorzystanie Sztucznej Inteligencji (AI), w tym dużych modeli językowych (*Large Language Model*, LLM) stojących za sukcesem inteligentnych chatbotów (np. ChatGPT).

Czym jest LLM? To rozbudowana sieć neuronowa uczona za pomocą ogromnej liczby tekstowych materiałów treningowych[26]. Architektura modelu i wielkość zbiorów tekstu, którymi jest on uczony, powodują, że ów model nabiera niezwykłych umiejętności w zakresie generowania i przetwarzania tekstów, prowadząc w ostatnim czasie do powszechnego wzrostu zainteresowania AI. Modele nie są oczywiście narzędziem, które zastąpi ludzi i wyręczy ich we wszystkich zadaniach wymagających przetwarzania i generowania tekstów. Mogą natomiast, właściwie użyte, być bardzo pomocne w pracy specjalistów z dziedzin humanistycznych, pozwalając wstępnie przygotowywać teksty, przetwarzać materiały źródłowe i wydobywać z nich oczekiwane informacje.

24 API jest interfejsem pozwalającym na komunikowanie się programów komputerowych.

25 *Federacja Bibliotek Cyfrowych*, fbc.pionier.net.pl [dostęp: 30.01.2024]

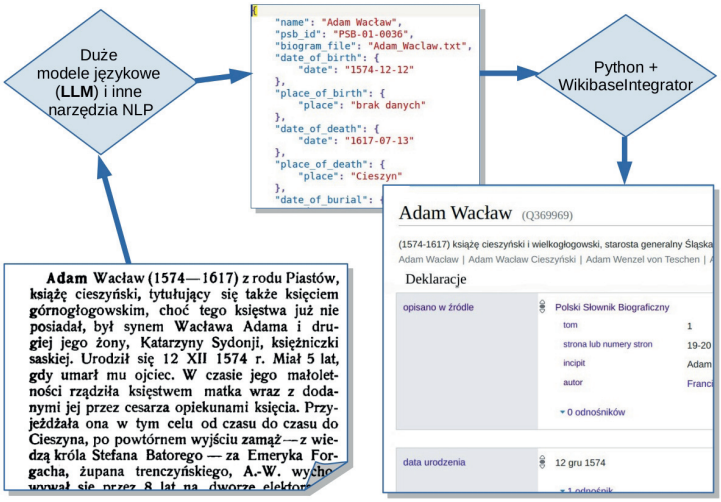
26 *Large language model*, en.wikipedia.org/wiki/Large_language_model [dostęp 20.09.2024]. Więcej na ich temat: M. Mamczur, *Czym jest i jak działa transformer (sieć neuronowa?)*, miroslawmamczur.pl/czym-jest-i-jak-dziala-transformer-siec-neuronowa [dostęp 30.11.2023]; S. Wolfram, *What is ChatGPT Doing… and Why Does It Work*, writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work [dostęp 30.11.2023]; A. Vaswani i in., *Attention Is All You Need*, arxiv.org/pdf/1706.03762v5.pdf [dostęp 30.11.2023]; J. Leike i in., *Training Language Models to Follow Instructions with Human Feedback*, arxiv.org/pdf/2203.02155.pdf [dostęp 30.11.2023]; *GPT-4 Technical Report*, arxiv.org/pdf/2303.08774.pdf [dostęp 30.11.2023].

LLM, jak każdy może się przekonać korzystając z ChatGPT[27] lub konkurencyjnych rozwiązań[28], potrafi całkiem naturalnie rozmawiać z człowiekiem, odpowiadać na jego pytania, przygotowywać materiały tekstowe[29] zgodnie z zadaną specyfikacją, ma też duże umiejętności translatorskie. Możliwość komunikacji w języku naturalnym z maszyną jest najszybciej zauważalną zmianą w świecie nowych technologii. Sam model językowy nie może być jednak traktowany jak baza wiedzy. Proces uczenia z wykorzystaniem milionów tekstów powoduje, że model dużo „wie". Sam model jest jednak przede wszystkim generatorem tekstu. Podczas procesu trenowania uczy się rozpoznawać wzorce językowe, struktury gramatyczne, kontekst słów i inne aspekty języka poprzez analizę ogromnej liczby dokumentów, od książek i artykułów po strony internetowe[30]. Gdy wytrenowany LLM otrzymuje tekst wejściowy (np. pytanie lub początek zdania), generuje odpowiedź bazując na swojej wiedzy i wzorcach językowych przewidując kolejne (najbardziej prawdopodobne) słowa w sekwencji, aż do osiągnięcia sensownego i spójnego zakończenia odpowiedzi. Generowanie kolejnego słowa wydaje się bardzo prostym mechanizmem, jednak pozwala na wykonywanie zaawansowanych zadań związanych z tekstami. „Wiedza" zaklęta w miliardach parametrów sieci neuronowej niekonieczne jest jednak wiedzą precyzyjną i aktualną. Materiał uczący duże modele zwykle składał się z tekstów w języku angielskim, a materiały w innych językach reprezentowane były w dużo mniejszym stopniu. Niepełność wiedzy modelu manifestuje się często w postaci tzw. halucynacji: model odpowiadając na pytania użytkownika potrafi zmyślać fakty, jest bowiem przede wszystkim sprawnym generatorem tekstu, a jego prawdziwość, a także wydźwięk to już zupełnie inna sprawa. Treści generowane przez model zależą od zawartości tekstów użytych do wytrenowania modelu, a te mogły zawierać fakty błędne, skażone preferencjami (np. politycznymi autorów) etc. Z takimi wadami modeli językowych można skutecznie walczyć i opracowywane są różne sposoby na poprawę jakości odpowiedzi modelu.

Model językowy można potraktować jako uniwersalne narzędzie przetwarzania języka naturalnego. Przydają się tu zarówno ogromne możliwości ekstrakcji wiedzy z tekstów przez model, rozumienie kontekstu, jak i możliwość sterowania działaniem modelu w języku naturalnym. Modele potrafią odpowiadać na pytania (na podstawie wewnętrznej lub dostarczonej w kontekście pytania wiedzy), tworzyć streszczenia, przygotowywać konspekty prac, wstępne wersje tekstów na podstawie dostarczonych notatek czy też tłumaczyć teksty na inne języki. Szczególnie dobre efekty osiągnięto w trakcie prac projektu DARIAH-PL[31], testując modele firmy OpenAI[32], np. model GPT-4[33], jako narzędzia do ekstrakcji informacji z tekstów biogramów z PSB. Ze zbioru biogramów przygotowywanych od blisko stu lat przez historyków wybrano reprezentatywną próbkę. Zadanie wydobycia z nich danych niezbędnych do wzbogacenia planowanej w ramach projektu bazy wiedzy opartej na systemie Wikibase byłoby bardzo praco- i czasochłonne, gdyby

27 *ChatGPT*, chat.openai.com/auth/login [dostęp 11.12.2023].
28 *Bard*, bard.google.com/chat [dostęp 11.12.2023].
29 Istnieją oczywiście modele zajmujące się obrazami, ich wytwarzaniem lub analizą, podobnie inne modele potrafią przetwarzać lub wytwarzać mowę, w tej pracy koncentrujemy się jednak na przetwarzaniu tekstów.
30 S. Wolfram, *What Is ChatGPT Doing*.
31 Projekt „Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce DARIAH-PL" finansowany w ramach Programu Operacyjnego Inteligentny Rozwój 2014-2020. Nr projektu: POIR.04.02.00-00-D006/20.
32 *OpenAI*, openai.com [dostęp 11.12.2023].
33 *GPT-4 Technical Report*.

oprzeć się tylko na pracy manualnej. Automatyczne przetwarzanie biogramów z użyciem mechanizmów wykorzystujących duże modele językowe pozwoliło na stworzenie czytelnych dla systemów komputerowych danych, nie tylko obejmujących podstawowe cechy osób (daty i miejsca urodzin czy śmierci), lecz także relacje rodzinne bohaterów biogramów, ich zawody, pełnione funkcje, związane z nimi instytucje czy też ważne postacie (niebędące krewnymi) występujące w ich życiorysie. Efekty nie były oczywiście bezbłędne, np. wyszukiwanie bliskich relacji rodzinnych (rodzice, rodzeństwo, dzieci) dawało wysoką skuteczność (90–99%), podczas gdy poprawność wyników dla dalszych stopni pokrewieństwa była odpowiednio niższa[34]. Wyniki te zostały jednak osiągnięte znacznie szybciej[35] i mniejszym kosztem niż w przypadku ręcznej pracy specjalistów, którzy bazując na przygotowanych w ten sposób danych mogą skupić się na ich weryfikacji i uzupełnianiu.



Ryc. 4. Proces automatycznego przetwarzania tekstu źródłowego na wiedzę w bazie WikiHum z użyciem narzędzi NLP (w tym dużych modeli językowych) (autor: Piotr Jaskulski)

Skoro wiedza modelu językowego nie jest idealna, jakie mamy sposoby by tę wiedzę wzbogacić? Obecny stan technologii pozwala na zastosowanie dwóch podejść: douczanie modelu, tzw. *fine-tuning*, zwykle polegający na dostrojeniu modelu przez trenowanie specjalnie przygotowaną serią przykładów, oraz wykorzystanie wiedzy zewnętrznej poprzez *retrieval-augmented generation* (RAG), czyli generowanie odpowiedzi na podstawie zapytania wzbogaconego odpowiednim kontekstem. Douczanie modelu jest z tych dwóch możliwości trudniejsze i dużo bardziej kosztowne (oczywiście to stan obecny, a technologia w dziedzinie AI zmienia się bardzo dynamicznie). Douczony model zapewne będzie udzielał odpowiedzi lepszej jakości w danej tematyce, ale nadal pozostanie problem halu-

34 *gtp_psb*, github.com/pjaskulski/gpt_psb [dostęp 11.12.2023]. Dokładny opis wyniku testów zostanie opublikowany w najbliższym czasie w formie osobnego artykułu: P. Jaskulski, T. Latos, M. Ryńca, A. Zapała, *Reliability of Large Language Models as a Tool for Knowledge Extraction from Biographical Dictionaries: the Case of the Polish Biographical Dictionary*, „Digital Scholarship in the Humanities" (w druku).

35 Model przetwarza jeden przeciętny biogram w 2–4 s, przy czym możliwe jest równoległe przetwarzanie paru biogramów.

cynacji. Istotnym elementem jest także czasochłonność przygotowywania materiałów do douczania, jak i konieczność częstego powtarzania tego procesu. Łatwiejsze do przeprowadzenia jest zbudowanie mechanizmu wyszukiwania informacji, które po dostarczeniu ich modelowi pozwolą na skuteczne udzielenie odpowiedzi na zadane pytanie.

Drugi ze wspomnianych sposobów wzbogacania wiedzy modelu językowego, mechanizmy typu RAG, wykorzystują najczęściej bazy wektorowe – jako źródło wiedzy, LLM – jako element interpretujący przekazaną wiedzę i przygotowujący odpowiedź na pytanie oraz prosty interfejs webowy – jako sposób komunikacji z użytkownikiem. Taki mechanizm, na przykład w formie chatbota (lub wspomaganej przez AI wyszukiwarki informacji), mógłby stać się ważnym narzędziem historyka. Kluczowym elementem jest skuteczne wyszukiwanie informacji, stosunkowo proste w przypadku danych zgromadzonych już w formie ustrukturyzowanej, np. w bazie danych, plikach XML czy grafie wiedzy. Wskazana jest tu co prawda znajomość jakiegoś języka zapytań (SQL, SPARQL, XQuery), ale od lat istnieją interfejsy graficzne znacząco ułatwiające przeszukiwanie i filtrowanie danych nawet bez takich kompetencji. W przypadku korpusów tekstów proste szukanie słów kluczowych w plikach, wyszukiwanie pełnotekstowe w repozytoriach dokumentów czy szukanie z użyciem wzorców (wyrażeń regularnych) jest znanym i skutecznym sposobem znajdowania informacji. Istnieje oczywiście wiele źródeł, które nie są jeszcze dostępne w formie zdigitalizowanej lub są zdigitalizowane w formie znacznie utrudniającej przeszukiwanie (pliki graficzne lub pdf bez warstwy tekstowej). Takie źródła wymagają wstępnego przygotowania, zanim będą mogły stać się przedmiotem dalszych prac. Wyzwaniem jest jednak wyszukiwanie semantyczne, biorące pod uwagę znaczenie szukanej informacji, nie tylko zgodność na poziomie znaków czy słów. Spopularyzowane wraz z pojawieniem się dużych modeli językowych bazy wektorowe pozwalają na przechowywanie matematycznych reprezentacji tekstów czy obrazów (tzw. osadzeń, *embeddings*[36]) i są często zewnętrznym źródłem informacji dla modelu językowego. Zapis wektorowy umożliwia przeszukiwanie bazy ze względu na podobieństwo tematyczne. Fragmenty tekstów np. zdanie mówiące, iż „król Zygmunt przybył do Krakowa", będzie matematycznie bardziej podobne do zdania „w Krakowie mieszkał król Zygmunt" niż do zdania mówiącego o czymś zupełnie innym, np. „wiosną drogi stawały się nieprzejezdne dla wozów". Posiadając duży zbiór tekstów można dzięki bazom wektorowym i osadzeniom stworzyć z takiego zbioru źródło przeszukiwalne tematycznie, z którego będzie mógł skorzystać duży model językowy. Rolą korzystającego z takiego systemu historyka jest zadanie jasno sformułowanego pytania w języku naturalnym. Rolą systemu informatycznego jest przekształcenie pytania do formy wektorów – osadzeń, zbadanie ich podobieństwa do zawartości bazy i przygotowanie najbardziej podobnych dokumentów / fragmentów dokumentów. Nie jest to jeszcze odpowiedź na pytanie, lecz półprodukt, z którego będzie w stanie skorzystać albo historyk lub historyczka, albo LLM, któremu zostanie zlecone przygotowania streszczonej odpowiedzi na

36 Ponieważ komputery rozumieją liczby, a nie tekst, informacje zapisane w tekście (słowie, zdaniu, akapicie czy całym dokumencie) zapisuje się w formie wektora liczb, wielowymiarowej reprezentacji liczbowej specyficznej dla danego tekstu. Wielkość wektorów różni się zależnie od technologii ich przygotowania, jeżeli użyjemy modelu text-ada-002 firmy OpenAI będą to wektory 1536 wymiarowe. Różnica między wektorami opisującymi dwa fragmenty tekstu przekłada się na ich podobieństwo tematyczne lub jego brak. Zob. *What are Embeddings?*, learn.microsoft.com/en-us/semantic-kernel/memories/embeddings [dostęp 11.12.2023].

bazie wyników wyszukiwania. Wyszukiwanie semantyczne nie musi być oczywiście jedynym źródłem informacji; dane można czerpać również z tradycyjnych baz danych czy grafowych baz wiedzy. Można sobie wyobrazić, jakim ułatwieniem w pracy historyka byłby system pozwalający na tak wszechstronne wyszukiwanie informacji, z możliwością przygotowywania streszczeń czy podsumowań przez AI, bez konieczności opanowywania technologii zapytań czy nauki interfejsu graficznego systemów wspomagających przeglądanie i filtrowanie danych[37].

Omawiając różne zalety i możliwości wykorzystania sztucznej inteligencji i modeli językowych, nie należy oczywiście zapominać o wadach i problemach tej technologii. AI popełnia błędy – mimo zaskakujących możliwości i efektów wykorzystania modele językowe nie są narzędziami bezbłędnymi. Czy to generując treść na podstawie danych, czy wyszukując informacje należy liczyć się z błędami zarówno dotyczącymi faktów (halucynacje modelu), jak i błędami językowymi. Wszystko, co AI stworzy, powinno więc podlegać weryfikacji przez człowieka lub przez inne systemy, również oparte na AI. AI mimo szybkiego rozwoju ciągle posiada pewne ograniczenia techniczne. Wielkość jednorazowo przetwarzanego tekstu (kontekstu, który przesyłamy do dużego modelu językowego), która jeszcze rok temu była poważnym limitem, jest dziś już bardzo duża (nawet setki tysięcy słów), nadal jednak dużo bardziej ograniczona jest długość zwracanego tekstu przygotowanego przez model AI. Kolejnym zagadnieniem jest problem praw autorskich: czy firmy udostępniające duże modele językowe w formie API lub chatu mogą przechowywać i wykorzystywać teksty wysyłane do przetwarzania? Stojąca za najpopularniejszymi dziś modelami GPT firma OpenAI gwarantuje prywatność udostępnianych modelom danych. Co jednak z efektami pracy modelu? Czy są one dziełem modelu, czy zlecającego mu pracę człowieka? Wewnętrzne przepisy OpenAI zapewniają, że to użytkownik jest właścicielem wszelkich wyników uzyskanych w ramach usług (czyli praw majątkowych) w zakresie dozwolonym przez prawo[38], należy jednak brać pod uwagę możliwe różnice w przepisach między różnymi krajami (USA a UE). Firma zapewnia również swoim użytkownikom komercyjnym rodzaj tarczy prawnej dotyczącej wykorzystania jej narzędzi. Poza aspektem majątkowym powstaje pytanie, kto jest właściwie autorem streszczenia przygotowanego przez model na podstawie wskazówek człowieka oraz czy i w jakim stopniu użycie w publikacji naukowej (biograficznej) tekstów generowanych przez AI jest dopuszczalne. Należy też pamiętać, że wykorzystywanie AI może być obecnie kosztowne, jednak rozwój modeli *open source* i inicjatywa polskiego środowiska naukowego (polski model PLLuM[39]) pozwalają wierzyć, że ten problem stanie się w przyszłości mniej istotny.

37 Technologię RAG już dziś wykorzystują nowe systemy przeznaczone do wspomagania naukowców np. produkt „Assistant by scite" oferowany przez scite.ai pozwala na zadawanie pytań w języku naturalnym, a wygenerowana odpowiedź zawiera odwołania do konkretnych publikacji naukowych związanych tematycznie z pytaniem, wyszukanych w bazie artykułów. System ten działa jednak głównie dla dziedzin przyrodniczych i ścisłych.

38 *Enterprise Privacy at OpenAI*, openai.com/enterprise-privacy [dostęp 11.12.2023].

39 *Powstanie pierwszy polski otwarty wielki model językowy – PLLuM*, opi.org.pl/powstanie-pierwszy-polski-otwarty-wielki-model-jezykowy-pllum [dostęp 11.12.2023].

Konkluzje

Podsumowując wykorzystanie LOD oraz LLM rysuje dość spójną wizję rozwoju prac nad hasłami biograficznymi. Badacz będzie analizował źródło i przetwarzał informacje z niego do formy bazodanowej. Baza danych będzie harmonizowana z centralnym zbiorem danych (bazą referencyjną) przy pomocy zmapowanego modelu danych i wzajemnego wykorzystywania identyfikatorów. Aplikacje (agregator, chatbot) będą przeszukiwały te dane i odpowiadały na pytania osoby zainteresowanej. Celem historyka/użytkownika będzie sformułowanie odpowiedniego zapytania, wydobycia informacji z zasobu i dalsza ich analiza na abstrakcyjnym, niedostępnym dla maszyny (w obecnej chwili) poziomie.

Rozwiązania te opierać się muszą na zdecentralizowanej społeczności uczonych. Każdy projekt będzie przygotowywał własne materiały. Ich harmonizacja będzie możliwa dzięki wykorzystaniu wypracowanych wspólnie oraz ogólnie zaakceptowanych standardów, zarówno w kwestii formatu, jak i modelu danych. Forma tekstowa biogramów będzie współistnieć z rozwiązaniami cyfrowymi, ale ze względu na jej mniejszą efektywności, będzie coraz bardziej marginalizowana.

Taka wizja, choć wydaje się futurystyczna, jest możliwa do zrealizowania, wymaga jednak zmiany sposobu myślenia badaczy. W szczególności chodzi o zaakceptowanie zmiany technologicznej wywołanej przez powszechny dostęp do komputerów, Internetu oraz narzędzi wykorzystujących sztuczną inteligencję. Zmiana podstawowego medium przekazu informacji, z druku na komputer, wymusza dostosowanie praktyki badawczej do nowych warunków. Główny cel przedsięwzięć biografistycznych, jakim zawsze było przedstawienie najważniejszych informacji o życiorysie opisywanych postaci w skrótowej formie, pozostaje bez zmian. Zmienia się jednak forma: z drukowanego tekstu do formy ustrukturyzowanej informacji. Jakość wdrożenia nowych rozwiązań zależy przede wszystkim od przyjęcia filozofii otwartej nauki i podjęcia prac teoretycznych dotyczących wykorzystywanych standardów. W tym przypadku szczególnie ważne jest rozwijanie ontologii dziedzinowych, które pozwalają najpełniej opisać minioną rzeczywistość. Zdefiniowania zjawisk historycznych za pomocą logiki formalnej jest nowym, ale fascynującym działaniem badawczych. Istnieją już co prawda ontologie wyższego rzędu (np. często wykorzystywany w humanistyce CIDOC CRM[40]), wielu zjawisk historycznych nie da się jednak za ich pomocą opisać. Bardzo ważnym zadaniem jest też rozbudowa baz referencyjnych i podjęcie wysiłku zmierzającego do harmonizacji z nimi istniejących i tworzonych zasobów biograficznych. Wydaje się, że ośrodkiem predestynowanym do podjęcia tego typu działań (a także testowania nowych form publikacji) jest Zakład Polskiego Słownika Biograficznego Instytutu Historii PAN. Zmiany te muszą jednak wynikać z szerokiego konsensusu badaczy, rozumiejących korzyści z nich wynikające.

Bibliografia

Enterprise Privacy at OpenAI, openai.com/enterprise-privacy [dostęp 11.12.2023].
FAIR Principles, www.go-fair.org/fair-principles/ [dostęp 30.01.2024].

40 *What is the CIDOC CRM?*, www.cidoc-crm.org [dostęp 31.01.2024].

HDL.NET® Information Services, handle.net [dostęp 11.12.2023].

Jaskulski P., Latos T., Ryńca M., Zapała A., *Reliability of Large Language Models as a Tool for Knowledge Extraction from Biographical Dictionaries: the Case of the Polish Biographical Dictionary*, „Digital Scholarship in the Humanities" (w druku).

Large language model, en.wikipedia.org/wiki/Large_language_model [dostęp 20.09.2024].

LinkedData, www.w3.org/wiki/LinkedData [dostęp 30.01.2024].

Mamczur M., *Czym jest i jak działa transformer (sieć neuronowa?)*, miroslawmamczur.pl/czym-jest-i-jak-dziala-transformer-siec-neuronowa/ [dostęp 30.11.2023].

Miles A., Matthews B., Beckett D., Brickley D., Wilson M., Rogers N., *SKOS: A Language to Describe Simple Knowledge Structures for the Web*, epubs.stfc.ac.uk/manifestation/685/SKOS-XTech2005.pdf [dostęp 20.09.2024].

GPT-4 Technical Report, arxiv.org/pdf/2303.08774.pdf [dostęp 30.11.2023].

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J., Lowe R., *Training Language Models to Follow Instructions with Human Feedback*, arxiv.org/pdf/2203.02155.pdf [dostęp 30.11.2023].

Powstanie pierwszy polski otwarty wielki model językowy – PLLuM, opi.org.pl/powstanie-pierwszy-polski-otwarty-wielki-model-jezykowy-pllum [dostęp 11.12.2023].

Resource Description Framework (RDF), www.w3.org/RDF [dostęp 11.12.2023].

Richards K., White R., Nicolson N. & Pyle R., *A Beginner's Guide to Persistent Identifiers*, www.gbif.org/document/80575/a-beginners-guide-to-persistent-identifiers [dostęp 24.09.2024], DOI 10.35035/mjgq-d052.

Stefanowicz B., *Informacja. Wiedza. Mądrość*, Warszawa 2023 (Biblioteka Wiadomości Statystycznych, t. 66).

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., *Attention Is All You Need*, arxiv.org/pdf/1706.03762v5.pdf [dostęp 30.11.2023].

Web Ontology Language (OWL), www.w3.org/OWL [dostęp 30.01.2024].

What are Embeddings?, learn.microsoft.com/en-us/semantic-kernel/memories/embeddings [dostęp 11.12.2023].

What is the CIDOC CRM?, www.cidoc-crm.org [dostęp 31.01.2024].

Wilkinson M., Dumontier M., Aalbersberg I. i in., *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, „Scientific Data" 2016, t. 3, 160018, DOI 10.1038/sdata.2016.18.

Wolfram S., *What is ChatGPT Doing… and Why Does It Work*, writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work [dostęp 30.11.2023].

Finansowanie

Podczas prac nad artykułem wykorzystano infrastrukturę Dariah.lab zbudowaną w ramach projektu „Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce DARIAH-PL" finansowany w ramach Programu Operacyjnego Inteligentny Rozwój 2014-2020. Nr projektu: POIR.04.02.00-00-D006/20.

lic. **Maria Guzik-Jureczka**, programistka w Pracowni Historii Cyfrowej Instytut Historii im. T. Manteuffla PAN.
e-mail: mguzik-jureczka@ihpan.edu.pl

mgr **Piotr Jaskulski**, programista w Pracowni Historii Cyfrowej Instytut Historii im. T. Manteuffla PAN.
e-mail: pjaskulski@ihpan.edu.pl

dr **Adam Zapała**, historyk, kierownik Pracowni Historii Cyfrowej Instytut Historii im. T. Manteuffla PAN.
e-mail: azapala@ihpan.edu.pl

Data zgłoszenia artykułu: 2 lutego 2024
Data przyjęcia do druku: 20 września 2024