dr Marcin Roszkowski
Katedra Informatologii,
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski
m.roszkowski@uw.edu.pl
ul. Nowy Świat 69
00-927 Warszawa

Integracja kartotek haseł wzorcowych nazw osobowych w semantycznej bazie wiedzy Wikidane

Abstrakt

Artykuł dotyczy problemu obecności kartotek haseł wzorcowych w środowisku sieciowym z punktu widzenia tzw. sieci danych (ang. Web of Data). Celem artykułu była próba pokazania sposobów funkcjonowania KWH poza kontekstem katalogu bibliotecznego na przykładzie integracji w semantycznej bazie wiedzy Wikidane. Celem badań było określenie zakresu wykorzystania bibliotecznych KHW jako źródeł informacji do kontroli form nazw osobowych w bazie Wikidane. W tym celu analizie poddano model danych w bazie Wikidane w którym zidentyfikowano elementy metadanych odpowiedzialne za realizację podstawowych funkcji kartotek haseł wzorcowych. W dalszej części przedmiotem analiz była ilościowa analiza wykorzystania siedmiu KHW w bazie wiedzy Wikidane oraz z uwzględnieniem polskojęzycznej Wikipedii.

Tworzenie i zarządzanie kartotekami haseł wzorcowych (KHW) są uznawane[1] za jedno z najbardziej kosztownych procesów w kontekście katalogowania zbiorów bibliotecznych. Jednak korzyści jakie płyną z ich stosowania przekładają się na wysoką jakość danych bibliograficznych udostępnianych użytkownikom i instytucjom. Kontrola form językowych dla haseł reprezentujących osoby, instytucje czy pojęcia odnoszących się do treści dokumentów to dążenie z jednej strony do zapewnienia spójność danych a z drugiej do zapewnienia pożądanego poziomu szczegółowości i kompletności wyszukiwania informacji w katalogach bibliotecznych i bibliograficznych bazach danych. Podstawową funkcją kartotek wzorcowych dla haseł formalnych jest więc jednoznaczne rozróżnianie (ang. *disambiguation*) elementów rekordu bibliograficznego poprzez stosowanie preferowanej nazwy językowej i grupowanie wokół danego hasła (ang. *collocation*) alternatywnych form językowych[2].

[1] B. Tillet, *Authority Control: State of the Art and New Perspectives*, „Cataloging and Classification Quarterly" 2004, nr, 3–4,s. 24.
[2] J. Niu, *Evolving Landscape in Name Authority Control*,„Cataloging & Classification Quarterly" 2013, nr 51(4), s. 405.

Pragmatyka tworzenia kartotek haseł wzorcowych zakłada korzystanie z wiarygodnych źródeł informacji w celu utworzenia lub przejęcia form językowych dla haseł. Są to m.in. encyklopedie ogólne i dziedzinowe, słowniki biograficzne czy też słowniki terminologiczne. Wraz z włączeniem się środowiska World Wide Web do obiegu informacji w nauce, również zasoby sieciowe są wykorzystywane jako źródła informacji przy tworzeniu haseł wzorcowych, co przejawia się zamieszczaniem adresów URL w rekordach wzorcowych w odpowiednich polach formatu MARC 21[3]. Tym samym katalogi biblioteczne oraz KHW stają się częścią globalnej sieci powiązań. Jednak tak rozumiana sieć opiera się przede wszystkim na hiperłączach, które wskazują na relacje między rekordem np. KHW a źródłem sieciowym w ramach semantyki formatu MARC 21, tzn. w ramach znaczenia danego pola, w którym użyto odnośnik URL. Taka interpretacja relacji hipertekstowych jest podstawą idei World Wide Web, czyli sieci dokumentów. Obecność KHW w środowisku sieciowym oznacza więc z jednej strony dostępność dla użytkowników za pośrednictwem przeglądarek internetowych a z drugiej realizowanie założeń WWW w postaci korzystania z hiperłączy do zewnętrznych źródeł informacji. Z drugiej strony, warto zadać pytanie, w jaki sposób biblioteczne kartoteki haseł wzorcowych są wykorzystywane poza kontekstem katalogu bibliotecznego oraz jaką funkcję pełnią w takich sytuacjach.

Przedmiotem artykułu jest zagadnienie wykorzystania bibliotecznych kartotek haseł wzorcowych w środowisku sieciowym poza kontekstem katalogu bibliotecznego w tzw. semantycznych bazach wiedzy. Rozważania te przedstawiono na przykładzie integracji wybranych KHW dla nazw osobowych w serwisie Wikidane, który spełnia wymogi definicyjne semantycznej bazy wiedzy. Z metodologicznego punktu widzenia w artykule opracowano charakterystykę modelu danych serwisu Wikidane, w którym zidentyfikowano elementy metadanych odpowiedzialne za realizację podstawowych funkcji kartotek haseł wzorcowych. W dalszej części przedmiotem analiz była ilościowa analiza wykorzystania siedmiu KHW w bazie wiedzy Wikidane oraz z uwzględnieniem polskojęzycznej Wikipedii.

Rozwój środowiska sieciowego zmierza w kierunku tzw. sieci danych (ang. *Web of Data*), gdzie relacje między jej zasobami są na jeszcze niższym poziomie niż <dokument>-<dokument>. Idea Sieci Semantycznej (ang. *Semantic Web*) opiera się na udostępnianiu zasobów informacji, które opisane są za pomocą formalnych języków reprezentacji wiedzy wykorzystujących standardy sieciowe oraz ustanawianiu relacji (również formalnie

[3] Np. stosowanie podpola #0

specyfikowanych) pomiędzy elementami strukturalnymi tych zbiorów. Wizja Sieci Semantycznej Tima Bernersa-Lee[4] jest obecnie realizowana w postaci ruchu Linked Data. „Jest to model publikowania danych w środowisku sieciowym, w którym wykorzystuje się określone standardy sieciowe i którego podstawą jest ustanawianie relacji między opisywanymi dokumentami, osobami, pojęciami, wydarzeniami itd. w celu optymalizacji procesów wyszukiwania i automatycznej eksploracji informacji”[5]. Główne założenia tej metodyki sprowadzają się do czterech wytycznych odnoszących się do sposobu reprezentacji i publikowania danych w środowisku sieciowym[6]:

1. Stosowanie standardu URI (ang. *Uniform Resource Identifier*) jako sposobu odwoływania się do elementów danych.
2. Wykorzystanie protokołu HTTP (ang. *Hypertext Transfer Protocol*) jako kanału komunikacji i przesyłania danych.
3. Reprezentacji danych z wykorzystaniem schematów metadanych i ontologii w formatach opartych na specyfikacji RDF (ang. *Resource Description Framework*).
4. Ustanawianiu relacji między elementami danych wewnątrz danej bazy oraz z zewnętrznymi źródłami informacji z wykorzystaniem mechanizmu identyfikacji URI.

Stosowanie identyfikatorów sieciowych w standardzie URI zapewnia stabilny system odwoływania się do opisywanych obiektów w środowisku sieciowym. Protokół HTTP to standaryzowany kanał komunikacji a RDF określa zasady reprezentacji informacji.

Kluczowe znaczenie dla stabilności odwołań w środowisku sieciowym ma stosowanie tzw. stałych identyfikatorów (ang. *persistent identifier*), których format zapisu jest zgodny ze standardem URI (ang. *Uniform Resource Identifier*). Poprzez termin ten rozumie się zapis jednoznacznie identyfikujący zasób cyfrowy (np. dokument, obiekt lub rekord bibliograficzny), który może być długoterminowo wykorzystywany na potrzeby wyszukiwania informacji. Nawet jeśli zasób zmieni swoją fizyczną lokalizację, jego identyfikator pozostanie niezmieniony a mechanizm jego interpretacji zapewni odpowiednie przekierowanie i dostęp do treści, które reprezentuje[7]. Istotna jest tutaj tzw. sieciowa transferowalność, tzn. możliwość wywołania danego identyfikatora w środowisku sieciowym i uzyskania informacji na temat

[4] T. Berners-Lee,J. Hendler, O. Lassila, *The Semantic Web*, „Scientific American”, 2001, nr 248(5), s. 34-43.
[5] M. Roszkowski, *Kartoteki nazw osobowych w środowisku sieciowym*, [online] „Biuletyn EBIB”, 2015, nr7, http://open.ebib.pl/ojs/index.php/ebib/article/view/380 [dostęp: 18.09.2016]
[6] T. Berners-Lee, *Linked Data - Design Issues*, [online], 2006,http://www.w3.org/DesignIssues/LinkedData.html [dostęp: 18.09.2016]
[7] Zob. *National Bibliographies in the Digital Age: Guidance and New Directions*, pod red. M. Zumer. Munchen, 2009.

danego zasobu cyfrowego, które zapisano w rekordzie jego identyfikatora. Oznacza to również zapewnienie odpowiedniej infrastruktury, która pozwoli na uzyskanie danych w formacie pożądanym przez osobę lub aplikację kontaktującą się z serwerem, do którego przekierowuje dany identyfikator (tzw. *content negotiation*). Przykładem stałych identyfikatorów jest system DOI (ang. Digital Object Identifier) oraz PURL (ang. Persistent URL). W przypadku DOI drugi warunek w odniesieniu do stałych identyfikatorów, czyli transferowalność sieciową warto zilustrować przykładem. W czasopiśmie „Toruńskie Studia Bibliologiczne" wszystkie artykuły otrzymują unikalny identyfikator DOI. Np. tekst pt. „Początki bibliografii lokalnej w Polsce. Józef Ignacy Kraszewski i jego bibliografia druków wileńskich"[8] opublikowany w Vol 8, No 2 (15) posiada identyfikator DOI: 10.12775/TSB.2015.017. Stabilność tego identyfikatora jest zapewniona poprzez fakt, że bazą DOI zarządza szereg agencji rejestrujących (DOI Registration Agency). W takiej postaci DOI nie jest jednak identyfikatorem sieciowym, jest on rozpoznawany tylko w bazach rejestrujących te identyfikatory[9]. Dopiero jego postać URI - http://dx.doi.org/10.12775/TSB.2015.017 zapewnia jego transferowalność w środowisku WWW. Oznacza to, że istnieje pewien mechanizm interpretujący ten adres (ang. *link resolver*), który po jego wywołaniu w sieci powoduje przekierowanie do lokalizacji publikacji na stronie WWW czasopisma, czyli pod jej adres URL (ang. *Uniform Resource Locator*)

RDF jest standardem reprezentacji wiedzy, który jest oficjalną rekomendacją Konsorcjum World Wide Web. Jest to deklaratywny model reprezentacji wiedzy, który opiera się na tzw. trójkach RDF pełniących funkcję elementarnych jednostek wypowiedzi w tym języku. Trójka RDF zbudowana jest z przedmiotu, który jest opisywany, predykatu wskazującego na opisywany atrybut lub relację oraz obiektu, który zawiera wartość opisywanej cechy, np.

obiekt1 → *nazwisko_i_imię* → *Adam Mickiewicz*

Oprócz formalnej składni, RDF zakłada odwoływanie się do każdego elementu trójki za pośrednictwem URI. W tym celu dla identyfikacji obiektu stosuje się stałe identyfikatory URI w danej bazie, dla predykatu - URI ze stosowanego schematu metadanych (np. Dublin Core Metadata Element Set) a dla wartości - zależnie od sytuacji ciąg znaków lub URI dla obiektu, np.:

http://viaf.org/viaf/64009368/ → *http://schema.org/name* → *Adam Mickiewicz*

[8] http://apcz.pl/czasopisma/index.php/TSB/article/view/TSB.2015.017
[9] Np. https://dx.doi.org/

Zastosowanie stałych identyfikatorów w postaci URI oraz formalnie specyfikowanego schematu metadanych (lub ontologii) w ramach modelu RDF pozwala na reprezentację informacji na poziomie faktów, składających się na opis kolekcji. Te, dzięki zastosowanym mechanizmom przetwarzania (np. automatycznego wnioskowania), tworzą tzw. graf wiedzy lub semantyczną bazę wiedzy. Mamy bowiem do czynienia z pewnym formalnym modelem fragmentu rzeczywistości, który reprezentuje schemat metadanych lub ontologia, zbiorem obiektów składających się na fragment rzeczywistości (np. dokumenty, osoby, miejsca, pojęcia) oraz zestawem reguł wnioskowania zależności między nimi oraz nowych faktów na podstawie danych już istniejących.

W tak naszkicowanej perspektywie interpretacja kartoteki haseł wzorcowych jako semantycznej bazy wiedzy zakłada konieczność opisu jej zawartości za pomocą schematu metadanych lub ontologii w ramach składni formalnych języków reprezentacji wiedzy (np. RDF) oraz odwoływanie się do elementów metadanych za pośrednictwem ich unikalnych identyfikatorów w postaci URI.

Ostatni punkt metodyki Linked Data zakłada odwoływanie się do zewnętrznych baz wiedzy poprzez ustanawianie relacji między obiektami (w ramach składni RDF i za pośrednictwem URI), które mogą być pożądane w celu dalszej eksploracji informacji.

Od sieci dokumentów do sieci danych - Wikidane

Wikidane[10] to społecznościowa baza wiedzy i centralna platforma zarządzania danymi na potrzeby Wikipedii i jej projektów siostrzanych (np. Wikiźródła i Wikicytaty)[11].U podstaw tego projektu leży założenie ruchu Wikimedia[12] – „Wyobraź sobie świat, w którym każda osoba ma dostęp do sumy ludzkiej wiedzy”[13] oraz jego interpretacja w kontekście rozwoju tzw. sieci danych (ang. *Web of Data*). W takim ujęciu Wikipedia powinna być traktowana również jako baza danych, która pozwala na dostęp zarówno użytkownikom jak i aplikacjom oraz zapewnia spójny mechanizm wyszukiwania informacji i ponownego użycia zebranych danych. Obecnie

[10] http://www.wikidata.org
[11] F. Erxleben, M. Günther, M. Krötzsch, J, Mendez, D. Vrandečić, *Introducing Wikidata to the Linked Data Web*, [w:] The Semantic Web – ISWC 2014. 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I, red. P. Mika et al., Berlin 2014, s. 50–65.
[12] http://wikimedia.org
[13] https://pl.wikimedia.org/

Wikipedia to ponad 30 milionów artykułów w 287 językach, z których ekstrakcja danych jest nie lada wyzwaniem. Sam proces wyszukiwania informacji w Wikipedii w większości przypadków ogranicza się do odnalezienia hasła encyklopedycznego. Natomiast bardziej złożone zapytania muszą być realizowane, albo poprzez manualne przeglądanie zawartości haseł encyklopedycznych, albo dzięki specjalnym kategoriom w Wikipedii, które agregują informacje z wielu haseł.[14] Dodatkowym ograniczeniem jest fakt, że informacje w Wikipedii są rozproszone na stronach w różnych jej wersjach językowych. Celem projektu Wikidane było więc z jednej strony zapewnienie nowego sposobu zarządzania danymi w Wikipedii a z drugiej udostępnienie mechanizmów do ich wyszukiwania i dalszego przetwarzania.

Projekt jest koordynowany przez Wikimedia i uruchomiono go w 2012 r. Pierwszy etap polegał na centralizacji wersji językowych Wikipedii za pośrednictwem repozytorium danych Wikidane, które zagregowałoby te informacje. Dla każdego unikalnego hasła w Wikipedii utworzono obiekt (ang. *item*), dla którego opracowano rekord zawierający informacje o artykułach w poszczególnych wersjach językowych w Wikipedii. W marcu 2012 r. wszystkie wersje językowe Wikipedii zostały zarejestrowane w bazie Wikidane. Na początku 2013 r. rozpoczął się drugi etap projektu, który polegał na ekstrakcji informacji ze stron zawierających hasła Wikipedii i zapisaniu ich zgodnie z przyjętym modelem i formatem danych (m.in. RDF) w bazie Wikidane. Przedmiotem transformacji były informacje ustrukturyzowane, tzn. treści zapisywane w szablonie *Infoboks*, który w interfejsie graficznym Wikipedii jest prezentowany w prawym panelu[15]. Obecnie baza Wikidane jest na bieżąco aktualizowana, udostępniono również mechanizm wyszukiwania i pobierania danych.

Projekt Wikidane charakteryzuje się kilkoma istotnymi cechami[16]:

1. Otwarta edycja: tak samo jak w Wikipedii, baza Wikidane jest otwarta na edycję i dodawanie nowych informacji przez użytkowników.
2. Kontrola społeczności: zarówno dane jaki sam schemat metadanych jest wynikiem współpracy użytkowników serwisu.
3. Pluralizm: serwis zezwala na współistnienie sprzecznych danych i zapewnia mechanizm organizacji tych informacji.

[14] Np. Lista osób, które urodziły się w 1952 r. - https://pl.wikipedia.org/wiki/Kategoria:Urodzeni_w_1952; lista filmów, które są pełnometrażowymi debiutami reżyserów - https://pl.wikipedia.org/wiki/Kategoria:Pe%C5%82nometra%C5%BCowe_debiuty_re%C5%BCyser%C3%B3w
[15] D. Vrandečić, M. Krötzsch. *Wikidata: A Free Collaborative Knowledgebase*„Communications of the ACM", 2014, nr 57(10), s. 78–85.
[16] Tamże, s. 78-79

4. Źródła informacji: baza zapewnia możliwość dodawania źródeł informacji, z których przejęto dane.
5. Wielojęzyczność: Wikidane oferuje odpowiedniki w wielu językach zarówno dla atrybutów jak i wartości danych. O ile w przypadku Wikipedii mamy do czynienia z wieloma jej wersjami językowymi, to w przypadku Wikidane istnieje jedna wspólna baza danych.
6. Łatwy dostęp: Serwis Wikidane zapewnia dostęp do swoich zasobów zarówno za pośrednictwem interfejsu graficznego użytkownika jak również interfejsów programistycznych (API, SPARQL endpoint), które pozwalają na wyszukiwanie i pobieranie danych w wielu formatach (np. RDF). Dodatkowo regularnie publikowane są kopie bazy danych w postaci plików do pobrania.
7. Permanentna ewolucja: baza danych jest na bieżąco aktualizowana dzięki pracy wikipedystów, użytkowników serwisu Wikidane oraz programistów. Większość rozwiązań technicznych jest wdrażana stopniowo jak szybko jest to możliwe.

Obecnie baza Wikidane rejestruje informacje o 20 075 348[17] obiektach z czego ponad 3 miliony to hasła osobowe.

Wikidane – model danych

Z punktu widzenia modelu danych Wikidane opierają się na reprezentacji informacji na temat obiektów z wykorzystaniem par atrybut-wartość. Zarówno obiekty jak i atrybuty są identyfikowane za pomocą formalnych wykładników w postaci stałych identyfikatorów zgodnych ze standardem URI. Identyfikatory obiektów posiadają przedrostek Q zaś atrybutów P. Np. obiekt *Warszawa* posiada identyfikator Q270 a atrybut *populacja* – P1082. Z punktu widzenia organizacji samego serwisu WWW Wikidane (tak jak Wikipedia) są zbudowane z systemu stron internetowych, które zawierają informacje w postaci ustrukturyzowanej. Dla każdego obiektu w bazie Wikidane istnieje strona internetowa, której adres URL zawiera jego formalny identyfikator (np. dla obiektu Q270 będzie to https://www.wikidata.org/wiki/Q270) która prezentuje informacje jego na temat[18].

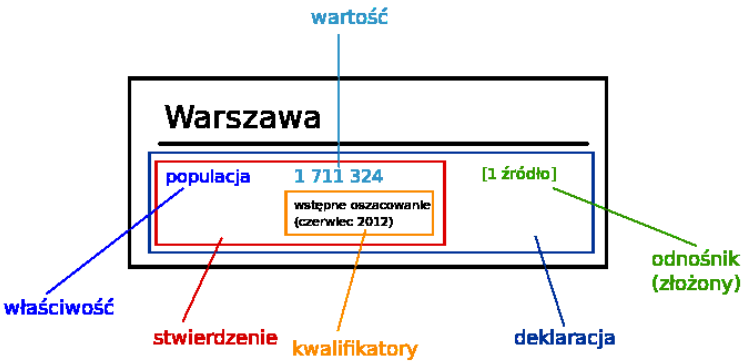
[17] Stan na 14.09.2016 r. na podstawie https://www.wikidata.org/wiki/Wikidata:Statistics/pl
[18] Obiekt *Warszawa* jest identyfikowany za pomocą URI w postaci http://www.wikidata.org/entity/Q270. Jeżeli z bazą kontaktuje się użytkownik za pośrednictwem przeglądarki zostanie on automatycznie przekierowany pod adres https://www.wikidata.org/wiki/Q270, gdzie informacje są wyświetlane w trybie graficznym.

Strona dla obiektu zawiera informacje zorganizowane w kilku sekcjach. Są to[19]:

1. Nazwa obiektu: np. Warszawa,
2. Krótka charakterystyka: np. stolica i największe miasto Polski,
3. Inne formy nazwy (tzw. alias): np. Warsaw, Warschau
4. Lista deklaracji na temat obiektu (dane ustrukturyzowane): np.
    - typ (P31) – miasto (Q515);
    - współrzędne geograficzne (P625)- 52°13'N, 21°2'E;
    - stolica dla (P1376) – Polska (Q36) ;
    - populacja (P1082) – 1 735 442.
5. Lista linków zewnętrznych (np. hasła encyklopedyczne w poszczególnych wersjach językowych Wikipedii) : np.
    - pl.wikipedia.org - https://pl.wikipedia.org/wiki/Warszawa
    - de.wikipedia.org - https://de.wikipedia.org/wiki/Warschau
    - pl.wikinews.org - https://pl.wikinews.org/wiki/Kategoria:Warszawa

Taka sama sytuacja ma miejsce w odniesieniu do informacji na temat atrybutów. W tym przypadku prezentowana jest jednak charakterystyka samego atrybutu (np. nazwa, atrybut nadrzędny) nie zaś obiekty, które zostały opisane za jego pomocą. Każde wystąpienie pary atrybut-wartość w odniesieniu do danego obiektu ma formę deklaracji (ang. *statement*), w której skład może wchodzić dodatkowo kwalifikator atrybutu (uszczegółowienie opisywanej własności) oraz informacja o źródle przejęcia danych (Rys. 1).Tym samym istnieje możliwość reprezentacji informacji o charakterze zmiennym, np. populacji Warszawy w odniesieniu do danego roku z podaniem źródła przejęcia tych danych. Taka sytuacja ma miejsce np. dla rekordu na temat Berlina (Q64)[20] gdzie informacje na temat populacji sięgają XIII w.

[19]F. Erxleben, M. Günther, M. Krötzsch, J, Mendez, D. Vrandečić, *Introducing Wikidata to the Linked Data Web*, [w:] The Semantic Web – ISWC 2014. 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I, red. P. Mika et al., Berlin 2014, s. 53.
[20] https://www.wikidata.org/wiki/Q64



Rysunek 1. Model danych w Wikidane. Źródło: https://pl.wikipedia.org/wiki/Wikidane

Wartości dla atrybutów są formalnie specyfikowane, tzn. model danych zakłada, że może to być m.in. ciąg znaków, współrzędne geograficzne, jednostki miar i wag czy też adresy URL obiektów zarówno wewnątrz bazy jak i do zewnętrznych źródeł informacji (np. atrybut stolica dla (P1376) odsyła w przykładzie do obiektu Polska (Q36) z bazy wiedzy Wikidane).

Wikidane centralizuje proces zarządzania danymi w Wikipedii rozumianej jako konglomerat jej wersji językowych. Agregacja danych z tych źródeł powoduje, że mamy dostęp do rożnych punktów widzenia na temat elementów rzeczywistości opisanych w krajowych Wikipediach. Przekłada się to na możliwość współistnienia również sprzecznych deklaracji opisujących dane obiekty, dzięki zapewnieniu mechanizmu identyfikacji źródła pochodzenia danych (ang. *provenance*). Wikidane czerpią również informacje z zewnętrznych źródeł. Założeniem projektu jest więc sytuacja, w której dla każdego faktu opisanego w bazie istnieje źródło jego pochodzenia, czy to w postaci wskazania na artykuł w danej wersji językowej Wikipedii, czy też poprzez odesłanie do zewnętrznej bazy danych. Należą do nich m.in. dokumenty, bazy danych statystycznych, bazy danych rządowych, kartoteki haseł wzorcowych, bazy bibliograficzne. Informacje o źródle pochodzenia danych zamieszcza się w bazie Wikidane za pomocą kwalifikatora uszczegóławiającego charakter źródła cytowanego. Służy do tego zestaw atrybutów, które zgrupowano pod nazwą „własność wskazująca źródła" (Q18608359) (Tab.1)

Tabela 1. Własność wskazująca źródła.

Identyfikator	Nazwa	Alias w j. polskim
http://www.wikidata.org/entity/P1343	described by source	opisano w źródle

http://www.wikidata.org/entity/P887	based on heuristic	oparty na heurystyce
http://www.wikidata.org/entity/P813	retrieved	data dostępu
http://www.wikidata.org/entity/P854	reference URL	URL źródła
http://www.wikidata.org/entity/P1683	quote	cytat
http://www.wikidata.org/entity/P143	imported from	pobrano z
http://www.wikidata.org/entity/P248	stated in	źródło
http://www.wikidata.org/entity/P1480	sourcing circumstances	status w źródle

Kartoteki haseł wzorcowych w Wikipedii

Koncepcja identyfikacji przedmiotów artykułów w Wikipedii za pośrednictwem m.in. kartotek haseł wzorcowych jest realizowana za pomocą szablonu o dosyć niefortunnej nazwy *Kontrola autorytatywna* (ang. *authority control*), którego zawartość jest widoczna w dolnej części strony zawierającej treść hasła encyklopedycznego. Za pomocą tego modułu użytkownicy Wikipedii mogą umieszczać w artykule dodatkowe identyfikatory dla hasła w Wikipedii, które przekierowują do kartotek haseł wzorowych lub rekordów bibliograficznych w katalogach bibliotecznych. Celem wprowadzenia tej funkcjonalności w Wikipedii było z jednej strony zapewnienie dodatkowej formy rozróżniania i identyfikacji nazw homonimicznych, co w nomenklaturze Wikipedii nazywa się „ujednoznacznianiem" a z drugiej świadome wykorzystanie dorobku bibliotek w zakresie tworzenia kartotek haseł wzorcowych na potrzeby środowiska sieciowego. Ustanawianie powiązań między Wikipedią a kartotekami haseł wzorcowych oraz katalogami bibliotecznymi w założeniu miało również dać podstawę programistom związanym z ruchem Wikimedia jak i środowiskiem bibliotecznym do tworzenia nowych narzędzi wykorzystujących potencjał zarówno Wikipedii jak i zasobów bibliotecznych.[21]

Moduł *Kontrola autorytatywna* pozwala obecnie na manualne umieszczanie identyfikatorów dla haseł w Wikipedii pochodzących z 34 zewnętrznych źródeł w postaci kartotek haseł wzorcowych oraz katalogów bibliotecznych. Są to m.in.:

- VIAF – Wirtualna Międzynarodowa Kartoteka Haseł Wzorcowych,
- LCCN -Numer Kontrolny Biblioteki Kongresu,
- ISNI - International Standard Name Identifier,

[21] https://en.wikipedia.org/wiki/Wikipedia:Authority_control

- ORCID – identyfikator ORCID,
- GND – kartoteka haseł wzorcowych Biblioteki Narodowej Niemiec - Gemeinsame Normdatei,
- SELIBR - kartoteka haseł wzorcowych szwedzkiego katalogu rozproszonego LIBRIS,
- BNF - kartoteka haseł wzorcowych Biblioteki Narodowej Francji,
- ULAN – kartoteka haseł osobowych Union List of Artist Names Getty Research Institute.
- WORLDCATID – identyfikator z katalogu WorldCat.

Z formalnego punktu widzenia polega to na wykorzystaniu odpowiedniej składni języka formatowania tekstu w artykułach (MediaWiki Markup Language), za pomocą której wskazuje się na skrót nazwy źródła danych oraz podaje identyfikator występującego tam hasła. Mechanizm zaimplementowany w module wyświetla te dane jako hiperłącza do bezpośredniej lokalizacji hasła w danej kartotece. Np. zawartość modułu dla hasła w Wikipedii *John Flanagan (pisarz)*[22] zawiera m.in.:

```
{{Kontrola autorytatywna |ISNI=0000 0001 1450 7820 |VIAF=85841451
|LCCN=n2009049211 |GND=132422573 BNF=155306391|SUDOC=150411979 …
}}
```

co w efekcie prowadzi do wyświetlenia w dolnej części strony wersji z hiperłączami (Rys.2).

Kontrola autorytatywna (osoba): ISNI: 0000 0001 1450 7820 | VIAF: 85841451
| LCCN: n2009049211 | GND: 132422573 | BnF: 155306391 | SUDOC: 150411979
| NKC: xx0081982 | NTA: 303496878 | Open Library: OL2660074A | WorldCat

Rysunek 2. Moduł Kontrola autorytatywna w Wikipedii.

W 2012 r. z inicjatywy wikipedystów zatrudnionych w OCLC i Biritsh Library (tzw. Wikipedian in Residence) uruchomiono projekt automatycznej identyfikacji haseł osobowych w Wikipedii w bazie VIAF. W efekcie powstało oprogramowanie VIAFbot[23], które na podstawie zastosowanych algorytmów mapowania umieściło identyfikatory VIAF dla ponad 250 tys. haseł osobowych w anglojęzycznej Wikipedii. Proces ten został przeprowadzony również dla niemieckojęzycznej Wikipedii z uwzględnieniem identyfikatorów GND. Efektywność zastosowanego algorytmu była dosyć wysoka. Odnotowane błędy stanowiły ok.

[22] https://pl.wikipedia.org/wiki/John_Flanagan_(pisarz)
[23] A. Kyrios, *VIAFbot and the Integration of Library Data on Wikipedia*, [online] „Code4Lib Journal”, 2013, nr 22, http: journal.code4lib.org/articles/8964 [dostęp: 19.09.2016].

10-15% przypadków. Uruchomiono również osobną stronę w Wikipedii, gdzie użytkownicy mogą zgłaszać dostrzeżone nieprawidłowości[24].

Dzięki integracji Wikipedii z bazą wiedzy Wikidane cały proces wypełniania treścią modułu *Kontrola autorytatywna* jest zautomatyzowany. Z punktu widzenia wikipedysty sprowadza on się wyłącznie do uruchomienia modułu[25] w artykule encyklopedycznym a jego zawartość (czyli identyfikatory) jest generowana na podstawie zasobów bazy Wikidane.

Kontrola danych w bazie Wikidane

Model Wikidane zakłada ustanawianie powiązań między obiektami tworzącymi bazę wiedzy a zewnętrznymi źródłami informacyjnymi, których celem jest szeroko pojęta normalizacja ich nazw. W tym celu w schemacie metadanych wyodrębniono klasę Q18614948 o nazwie *Właściwość w Wikidanych do kontroli autorytatywnej* (ang. *Wikidata property for authority control*). Jej celem jest zgrupowanie własności, które mają służyć do specyfikowania relacji między obiektem w bazie Wikidane a jego odpowiednikiem m.in. w kartotekach haseł wzorcowych, katalogach bibliotecznych, dziedzinowych bazach danych oraz dokumentach normalizacyjnych. W ramach tej klasy wyodrębniono dodatkowe 10 podklas (Tab. 2), w celu uporządkowania źródeł informacji pełniących funkcję punktu odniesienia w procesie kontroli danych.

Tabela 2.Specyfikacja klasy Właściwość w Wikidanych do kontroli autorytatywnej (Q18614948)

Identyfikator	Nazwa	Alias w j. polskim	Liczba źródeł
wd: Q18614948	Wikidata property for authority control	Właściwość w Wikidanych do kontroli autorytatywnej	148
wd:Q18618628	Wikidata property for cultural heritage identification	Właściwość w Wikidanych do identyfikacji dziedzictwa kulturowego	62
wd:Q19595382	Wikidata property for authority control for people	Właściwość w Wikidanych do kontroli autorytatywnej osób	324
wd:Q19829908	Wikidata property for authority control for places	Właściwość w Wikidanych do kontroli autorytatywnej miejsc	119

[24] https://en.wikipedia.org/wiki/Wikipedia:VIAF/errors
[25] Polega to na edycji strony dla danego hasła encyklopedycznego i umieszczeniu w dolnej jej części deklaracji {{Kontrola autorytatywna}} o pustej treści.

wd:Q19833377	Wikidata property for authority control for works	Właściwość w Wikidanych do kontroli autorytatywnej prac twórczych	132
wd:Q19833835	Wikidata property for authority control for substances	Właściwość w Wikidanych do kontroli autorytatywnej substancji	38
wd:Q21745557	Wikidata property for authority control for organisations	Właściwość w Wikidanych do kontroli autorytatywnej organizacji	56
wd:Q22964274	Wikidata property for identification in the film industry	Właściwość w Wikidanych do identyfikacji w przemyśle filmowym	41
wd:Q24075706	Wikidata property for authority control, with reciprocal use of Wikidata	Właściwość w Wikidanych do kontroli autorytatywnej z wzajemnym wykorzystaniem Wikidanych	15
wd:Q24575337	Wikidata property for authority control for events	Właściwość w Wikidanych do kontroli autorytatywnej wydarzeń	4
wd:Q26696664	Wikidata property for identifiers in product and service registers	Właściwości w Wikidanych do kontroli autorytatywnej produktów i usług	1

W bazie Wikidane stosuje się łącznie 940 źródeł informacji pełniących funkcję kartotek wzorcowych. Źródła danych obejmujące różne kategorie haseł (148) zgrupowano w klasie Q18614948. Pozostałe kartoteki i bazy danych przyporządkowano do klas reprezentujących poszczególne kategorie haseł. Z danych przedstawionych w Tab. 1 wynika, że najwięcej źródeł informacji w celu szeroko rozumianej kontroli haseł odnosi się do nazw osobowych (34%), wytworów aktywności intelektualnej i artystycznej, czyli dzieł (14%) oraz nazw miejscowych (12%). Wśród kartotek o charakterze ogólnym, które zawierają hasła osobowe znalazły się kartoteki haseł wzorcowych bibliotek narodowych, np.:

- identyfikator Bnf ID: Biblioteki Narodowej Francji (http://www.wikidata.org/entity/P268),
- identyfikator GND ID: Gemeinsame Normdatei Biblioteki Narodowej Niemiec (http://www.wikidata.org/entity/P227),
- identyfikator BNE ID: Biblioteki Narodowej Hiszpanii (http://www.wikidata.org/entity/P950)
- identyfikator LC ID: kartoteki Biblioteki Kongresu Library of Congress Name Authority File (http://www.wikidata.org/entity/P244)

W rekordach haseł osobowych w bazie Wikidane i tym samym Wikipedii wykorzystuje się również identyfikator ISNI - International Standard Name Identifier (http://www.wikidata.org/entity/P213) oraz VIAF (http://www.wikidata.org/entity/P214).

Wybrane kartoteki haseł wzorcowych dla nazw osobowych w bazie Wikidane

Celem badania było uzyskanie odpowiedzi na pytanie o zakres wykorzystania wybranych kartotek haseł wzorcowych dla nazw osobowych w bazie Wikidane. Jako przedmiot badań wybrano następujące KHW:

- kartoteka VIAF (VIAF ID),
- kartoteka haseł wzorcowych Biblioteki Kongresu (LCNAF ID),
- kartoteka Gemeinsame Normdatei Biblioteki Narodowej Niemiec (GND ID),
- kartoteka haseł wzorcowych Biblioteki Narodowej Francji (BnF ID),
- kartoteka haseł wzorcowych Biblioteki Narodowej Hiszpanii (BNE ID),
- kartoteka haseł wzorcowych NUKAT (NUKAT ID),
- kartoteka haseł wzorcowych Biblioteki Narodowej (NLP ID).

Celem badania było określenie liczby haseł osobowych w bazie Wikidane, w których rekordach występują identyfikatory ze wspomnianych kartotek z dodatkowym uwzględnieniem polskojęzycznej Wikipedii. W celu pozyskania danych wykorzystano interfejs programistyczny Wikidanych w postaci tzw. SPARQL Endpoint (https://query.wikidata.org/). Ten rodzaj trybu dostępu do semantycznych baz wiedzy polega na samodzielnej konstrukcji zapytania w języku SPARQL. Badania przeprowadzono 12.09.2016 r.

Baza Wikidane zawiera informacje na temat 3 247 156 osób. Informację tę uzyskano zliczając elementy bazy wiedzy, które w rekordach metadanych posiadają zapis *jest instancją klasy człowiek*, czyli posiadają atrybut P31 (jest instancją - ang. *instance of*) o wartości Q5 (człowiek - ang. *human*).

Tabela 3 zawiera dane dotyczące liczby haseł w bazie Wikidane, w których rekordach występują identyfikatory z kartotek haseł wzorcowych, które uwzględniono w badaniu. Średnio co czwarte hasło osobowe zawiera identyfikator VIAF. Udział pozostałych kartotek z wyjątkiem NUKAT oscyluje w granicach 10%. W przypadku KHW NUKAT i Biblioteki Narodowej, kwestia implementacji w Wikidanych jest bardziej złożona i wymaga komentarza.

Tabela 3. Hasła osobowe z identyfikatorami z siedmiu KHW

KHW	Identyfikator KHW	Hasła osobowe z identyfikatorem KHW	Hasła osobowe z identyfikatorem KHW (%)
VIAF	P214	839515	25,9
LC ID	P244	308123	9,5

GND ID	P227	349911	10,8
BnF ID	P268	292071	9,0
BNE ID	P950	30202	0,9
NUKAT ID	P1207	21657	0,7
NLP ID	P1695	7460	0,2

Biblioteczne kartoteki haseł wzorcowych, które występują w bazie Wikidanych jako źródła informacji w procesie ujednoznaczniania nazw, są wykorzystywane do tego procesu poprzez ustanawianie formalnych powiązań między opisywanym rekordem a rekordem odpowiedniego hasła wzorcowego. Relacja ta wiąże obiekt w bazie Wikidane z obiektem KHW za pośrednictwem ich unikalnych identyfikatorów. O ile baza Wikidane zapewnia ujednolicony system stałych identyfikatorów za pomocą standardu URI, to nie wszystkie biblioteczne KHW oferują taki tryb dostępu do swoich zasobów. W grupie KHW, które były przedmiotem badania taka sytuacja występuje w odniesieniu do KHW NUKAT.W tym przypadku rekordy wzorcowe w bazie NUKAT posiadają dane identyfikujące zapisywane w postaci numeru systemowego rekordu (pole 001 MARC 21) oraz numeru kontrolnego rekordu (pole 010 MARC 21). Np. rekord wzorcowy dla hasła osobowego *Miłosz, Czesław (1911-2004)* w bazie NUKAT zawiera pola:

01 vtls002124179

010 $a n93126971

Z punktu widzenia obecności KHW NUKAT w środowisku sieciowym, wspomniany rekord wzorcowy dostępny jest pod adresem http://katalog.nukat.edu.pl/lib/authority?id=875823. Jest to identyfikator w postaci adresu URL, który zawiera w treści zapytani do bazy NUKAT o wyświetlenie informacji o obiekcie, który posiada wewnętrzny identyfikator 875823. W bazie Wikidane rekord dla tego hasła (Q45970) w polu *identyfikator NUKAT* zawiera zapis: n93126971, co przekierowuje pod adres: http://viaf.org/processed/NUKAT|n93126971. Sytuacja ta wynika z faktu, że NUKAT współpracuje z VIAF w zakresie tworzenia tej międzynarodowej kartoteki haseł wzorcowych i informacja o rekordach przesłanych przez instytucje współpracujące jest dostępna w tzw. klastrach VIAF[26]. Tym samym integracja NUKAT jako źródła informacji dla haseł w bazie Wikidane odbywa się za pośrednictwem

[26] Zob. M. Roszkowski, *Kartoteka haseł wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF*. [w:] *Bibliografia – teoria, praktyka, dydaktyka*, red. J. Woźniak-Kasperek, J. Franke. Warszawa, 2016, s. 203–222.

VIAF. Taki sposób identyfikacji haseł wzorcowych z bibliotecznych KHW w Wikidanych dotyczy również np. Biblioteki Watykańskiej (BAV ID).

W przypadku wykorzystania w bazie Wikidane na potrzeby identyfikacji nazw osobowych kartoteki haseł wzorcowych Biblioteki Narodowej sytuacja wygląda nieco inaczej. W rekordzie w Wikidanych w polu NLP ID występuje wartość A10856754, która przekierowuje do Kartoteki Haseł Wzorcowych Formalnych BN pod adres:

http://mak.bn.org.pl/cgi-bin/KHW/makwww.exe?BM=01&IM=03&NU=01&WI=A10856754

Tutaj w rekordzie wzorcowym w polu 001 jako numer systemowy występuje wartość a10856778. Z punktu widzenia identyfikacji hasła osobowego, adres ten pełni taką samą funkcję jak link do rekordu wzorcowego w bazie NUKAT (czyli jest to URL). Chociaż Biblioteka Narodowa dostarcza rekordy wzorcowe do VIAF, to w tym przypadku jej integracja z bazą Wikidane odbywa się za pośrednictwem bezpośredniego łącza do strony internetowej zawierającej rekord wzorcowy (URL).

Kolejnym etapem badań była próba udzielenia odpowiedzi na pytanie o liczbę haseł osobowych w polskojęzycznej Wikipedii, w których występują identyfikatory haseł wzorcowych z siedmiu wspomnianych kartotek haseł wzorcowych. Dane uzyskane za pośrednictwem Wikidanych prezentuje tabela 4.

Tabela 4. Hasła osobowe z polskojęzycznej Wikipedii z identyfikatorami KHW

KHW	Identyfikator KHW	Hasła osobowe w pl.wikipedia.org z identyfikatorem KHW	Hasła osobowe pl.wikipedia.org z identyfikatorem KHW (%)
VIAF	P214	106445	36,1
LC ID	P244	58857	20,0
GND ID	P227	59904	20,3
BnF ID	P268	49370	16,8
BNE ID	P950	10366	3,5
NUKAT ID	P1207	7300	2,5
NLP ID	P1695	2349	0,8

Polskojęzyczna Wikipedia według danych z Wikidane rejestruje 294 627 haseł osobowych. Ponad 36% haseł osobowych posiada identyfikator VIAF a średnio co piąte identyfikatory Biblioteki Kongresu i Biblioteki Narodowej Niemiec. Udział KHW NUKAT i Biblioteki Narodowej jest niewielki. Pierwszym przypadku to nieco ponad 2,5% haseł a w drugim poniżej jednego procenta przypadków. Co ciekawe, tylko w 18% przypadków są to hasła, które jednocześnie zawierają identyfikatory NUKAT ID i BN ID.

Zastanawiać może również duża dysproporcja między liczbą haseł osobowych w polskiej Wikipedii, które posiadają identyfikator VIAF oraz liczbą identyfikatorów NUKAT i BN. Wśród tej puli haseł są np.:

- George W. Bush (Q207),
- Benedykt XVI (Q2494),
- Paul Otlet (Q1868),
- Jacques Chirac (Q2105),
- Konrad Adenauer (Q2492).

Warto odnotować, że w rekordach VIAF dla tych haseł są informacje również z NUKAT. Jedną z możliwych przyczyn, chociaż raczej o charakterze przypuszczenia, może być automatyzacja procesu dodawania identyfikatorów przede wszystkim z KHW NUKAT za pośrednictwem bazy VIAF oraz czas, w którym on miała miejsce. Można przypuszczać, iż w momencie pobierania danych z VIAF w tej bazie nie istniały jeszcze rekordy przesłane przez NUKAT, stąd brak tych danych w Wikidanych.

Z przedstawionych danych badawczych wynika, że największy udział w procesie kontroli danych dla haseł osobowych w bazie Wikidane ma kartoteka VIAF. Niestety zakres powiązań haseł osobowych z rekordami wzorcowymi obejmuje tylko jedną czwartą bazy Wikidane. W przypadku Polski, czyli udziału NUKAT i Biblioteki Narodowej zakres stosowania tych źródeł jest niewielki. Pomimo udziału tych dwóch ośrodków w projekcie VIAF, pojawia się problem konieczności aktualizacji tych danych.

Zakończenie

Integracja bibliotecznych kartotek haseł wzorcowych w semantycznych bazach danych, czy też szerzej – w projektach realizujących założenia sieci semantycznej wymaga ich przygotowania pod względem formalnym, tzn. spełnienia wymogów usługi sieciowej[27] zgodnie z obowiązującymi standardami sieciowymi. Dotyczy to zarówno reprezentacji informacji za pomocą formalnie specyfikowanych schematów metadanych jak i stosowania stałych identyfikatorów w standardzie URI. Pozwoli to zaistnieć tego rodzaju wartościowym źródłom

[27] Zob. np. M. Roszkowski, *Kartoteka haseł wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF*. [w:] *Bibliografia – teoria, praktyka, dydaktyka*, red. J. Woźniak-Kasperek, J. Franke. Warszawa 2016, s. 203–222

informacji, które przez lata prowadzenia przez biblioteki osiągnęły wysoki poziom jakości prezentowanych informacji, w środowisku sieciowym jako pełnoprawny element tzw. sieci danych. Kartoteki haseł wzorcowych mogą spełnić istotną funkcję w kontekście rozwoju sieci semantycznej. O ile w środowisku sieciowym istnieje wiele schematów metadanych i ontologii o charakterze ogólnym albo opracowanych na potrzeby dziedzinowych baz danych, to wydaje się, że to czego potrzebuje to środowisko to wysokiej jakości zbiory słownictwa kontrolowanego, które zapewnią pożądany poziom spójności danych.