

MARCIN ROSZKOWSKI

(Wydział Dziennikarstwa, Informacji i Bibliologii, Uniwersytet Warszawski, Warszawa)

INTEGRACJA KARTOTEK HASEŁ WZORCOWYCH NAZW OSOBOWYCH W SEMANTYCZNEJ BAZIE WIEDZY WIKIDANE

Tworzenie kartotek haseł wzorcowych (khw) i zarządzanie nimi to jeden z najkosztowniejszych procesów w kontekście katalogowania zbiorów bibliotecznych¹. Korzyści, które płyną z ich stosowania, przekładają się na wysoką jakość danych bibliograficznych udostępnianych użytkownikom i instytucjom. Kontrola form językowych haseł reprezentujących osoby, instytucje czy pojęcia odnoszących się do treści dokumentów to dążenie z jednej strony do zapewnienia spójności danych, a z drugiej strony do zapewnienia pożądanego poziomu szczegółowości i kompletności wyszukiwania informacji zawartych 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 językowych².

Pragmatyka tworzenia kartotek haseł wzorcowych zakłada korzystanie z wiarygodnych źródeł informacji w celu utworzenia lub przejścia form językowych dla haseł. Są to między innymi 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 (*Uniform Resource Locator*) w rekordach wzorcowych w odpowiednich polach formatu MARC 21³.

¹ B. Tillet, *Authority Control. State of the Art and New Perspectives*, „Cataloging and Classification Quarterly” 2004, nr 3–4, p. 24.

² J. Ni u, *Evolving Landscape in Name Authority Control*, „Cataloging & Classification Quarterly” 2013, nr 51(4), p. 405.

³ Na przykład stosowanie podpoła #0.

Tym samym katalogi biblioteczne oraz khw stają się częścią globalnej sieci powiązań. Tak rozumiana sieć opiera się przede wszystkim na hiperłączach, które wskazują na relacje między rekordem na przykład khw a źródłem sieciowym w ramach semantyki formatu MARC 21, czyli w ramach znaczenia danego pola, w którym użyto odnośnika 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. Warto też 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 tak zwanych 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 przeprowadzono ilościową analizę wykorzystania siedmiu khw w bazie wiedzy Wikidane z uwzględnieniem polskojęzycznej Wikipedii.

Rozwój środowiska sieciowego zmierza w kierunku tak zwanej 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 na ustanawianiu relacji (również formalnie specyfikowanych) pomiędzy elementami strukturalnymi tych zbiorów. Wizja Sieci Semantycznej Tima Bernersa-Lee⁴ 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 sieciowym⁶, a mianowicie:

⁴ T. Berners - Lee, J. Hendler, O. Lassila, *The Semantic Web*, „Scientific American” 2001, nr 248(5), p. 34–43.

⁵ M. Roszkowski, *Kartoteki nazw osobowych w środowisku sieciowym*, „Biuletyn EBIB” 2015, nr 7, [online] <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].

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. Reprezentacja danych z wykorzystaniem schematów metadanych i ontologii w formatach opartych na specyfikacji RDF (ang. *Resource Description Framework*).
4. Ustanawianie relacji między elementami danych wewnątrz danej bazy oraz z zewnętrznymi źródłami informacji przy wykorzystaniu 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 tak zwanych stałych identyfikatorów (ang. *persistent identifier*), których format zapisu jest zgodny ze standardem URI. Poprzez termin ten rozumie się zapis jednoznacznie identyfikujący zasób cyfrowy (na przykład dokument, obiekt lub rekord bibliograficzny), który to zapis może być długoterminowo wykorzystywany na potrzeby wyszukiwania informacji. Nawet jeśli zasób zmieni swoją fizyczną lokalizację, jego identyfikator pozostanie niezmienny, a mechanizm jego interpretacji zapewni odpowiednie przekierowanie i dostęp do treści, które reprezentuje⁷. Istotna jest tutaj tak zwana siecowa transferowalność, to znaczy możliwość wywołania danego identyfikatora w środowisku sieciowym i uzyskania informacji na temat danego zasobu cyfrowego zapisanych w rekordzie jego identyfikatora. Oznacza to również zapewnienie odpowiedniej infrastruktury, pozwalającej uzyskać dane w formacie pożądanym przez osobę lub aplikację kontaktującą się z serwerem, do którego przekierowuje określony 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 siecowa warto zilustrować przykładem. W czasopiśmie „Toruńskie Studia Bibliologiczne” wszystkie artykuły otrzymują unikatowy identyfikator DOI. Na przykład tekst zatytułowany *Początki bibliografii lokalnej w Polsce. Józef Ignacy Kraszewski i jego bibliografia druków wileńskich*⁸, opublikowany w R. 8, nr 2 (15), ma identyfikator DOI: 10.12775/TSB.2015.017. Stabilność tego identyfikatora jest zapewniona dzięki temu, 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 identyfikatory⁹. Dopiero jego postać URI – <http://dx.doi.org/10.12775/TSB.2015.017> zapewnia jego transferowalność w środowisku WWW. Oznacza to, że istnieje pewien me-

⁷ Zob. *National Bibliographies in the Digital Age. Guidance and New Directions*, red. M. Z u m e r, München 2009.

⁸ Zob. [online] <http://apcz.pl/czasopisma/index.php/TSB/article/view/TSB.2015.017> [dostęp: 2.11.2017].

⁹ Na przykład [online] <https://dx.doi.org/> [dostęp: 2.11.2017].

chanizm 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.

RDF jest standardem reprezentacji wiedzy, który jest oficjalną rekomendacją Konsorcjum World Wide Web. Jest to deklaracyjny model reprezentacji wiedzy, który opiera się na tak zwanych 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 zawierającego wartość opisywanej cechy, na przykład:

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 (na przykład Dublin Core Metadata Element Set), a dla wartości – zależnie od sytuacji ciąg znaków lub URI dla obiektu, na przykład:

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

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. Dzięki zastosowanym mechanizmom przetwarzania (na przykład automatycznego wnioskowania) tworzą one tak zwany 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 (na przykład 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 (na przykład RDF) oraz odwoływanie się do elementów metadanych za pośrednictwem ich unikatowych 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

Wikidane¹⁰ to społecznościowa baza wiedzy i centralna platforma zarządzania danymi na potrzeby Wikipedii i jej projektów siostrzanych (na przykład Wikiźródła i Wi-

¹⁰ Zob. [online] <http://www.wikidata.org> [dostęp: 2.11.2017].

kicytaty)¹¹. U podstaw tego projektu leży założenie ruchu Wikimedia¹² – „Wyobraź sobie świat, w którym każda osoba ma dostęp do sumy ludzkiej wiedzy”¹³ oraz jego interpretacja w kontekście rozwoju tak zwanych 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 jednocześnie spójny mechanizm wyszukiwania informacji i ponownego użycia zebranych danych. Obecnie Wikipedia to ponad 30 milionów artykułów w 287 językach, z których ekstrakcja danych jest nie lada wyzwaniem. Sam proces wyszukiwania informacji w Wikipedii najczęściej ogranicza się do odnalezienia hasła encyklopedycznego. Natomiast bardziej złożone zapytania muszą być realizowane poprzez manualne przeglądanie zawartości haseł encyklopedycznych albo dzięki specjalnym kategoriom w Wikipedii, które agregują informacje z wielu haseł¹⁴. 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 roku. Pierwszy etap polegał na centralizacji wersji językowych Wikipedii, które za pośrednictwem repozytorium danych Wikidane zagregowałyby te informacje. Dla każdego unikatowego hasła w Wikipedii utworzono obiekt (ang. *item*) i opracowano dla którego rekord zawierający informacje o artykułach w poszczególnych wersjach językowych w Wikipedii. W marcu 2012 roku wszystkie wersje językowe Wikipedii zostały zarejestrowane w bazie Wikidane. Na początku 2013 roku 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 (miedzy innymi RDF) w bazie Wikidane. Przedmiotem transformacji były informacje ustrukturyzowane, to znaczy treści zapisywane w szablonie *Infoboks*, który w interfejsie graficznym Wikipedii jest prezentowany w prawym panelu¹⁵. Obecnie baza Wikidane jest na bieżąco aktualizowana, udostępniło również mechanizm wyszukiwania i pobierania danych.

Projekt Wikidane charakteryzuje się kilkoma istotnymi cechami¹⁶:

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 1, red. P. Mika et al., Berlin 2014, s. 50–65.

¹² Zob. [online] <http://wikimedia.org> [dostęp: 2.11.2017].

¹³ Tamże.

¹⁴ Na przykład lista osób, które urodziły się w 1952 roku – [online] https://pl.wikipedia.org/wiki/Kategoria:Urodzeni_w_1952 [dostęp: 2.11.2017]; lista filmów, które są pełnometrażowymi debiutami reżyserów – [online] https://pl.wikipedia.org/wiki/Kategoria:Pe%C5%82nometra%C5%BCowe_debiuty_re%C5%BCyser%C3%B3w [dostęp: 2.11.2017].

¹⁵ D. Vrandečić, M. Krötzsch, *Wikidata. A Free Collaborative Knowledgebase*, „Communications of the ACM” 2014, nr 57(10), p. 78–85.

¹⁶ Tamże, p. 78–79.

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, jak i sam schemat metadanych jest wynikiem współpracy użytkowników serwisu.

3. Pluralizm: serwis zezwala na współistnienie sprzecznych danych i zapewnia mechanizm organizacji tych informacji.

4. Źródła informacji: baza zapewnia możliwość dodawania źródeł informacji, z których przejęto dane.

5. Wielojęzyczność: baza 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 (na przykład 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 tak szybko, jak jest to możliwe.

Obecnie baza Wikidane rejestruje informacje o 20 075 348¹⁷ 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 identyfikatory atrybutów przedrostek P, na przykład 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 (na przykład dla obiektu Q270 będzie to <https://www.wikidata.org/wiki/Q270>), która prezentuje informacje jego na temat¹⁸.

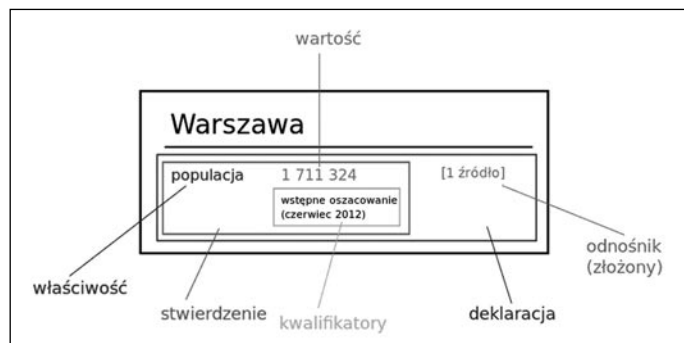
¹⁷ Stan na 14.09.2016 roku na podstawie <https://www.wikidata.org/wiki/Wikidata:Statistics/pl>.

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 to¹⁹:

1. Nazwa obiektu: na przykład Warszawa,
2. Krótka charakterystyka: na przykład stolica i największe miasto Polski,
3. Inne formy nazwy (tak zwany alias): na przykład Warsaw, Warschau
4. Lista deklaracji na temat obiektu (dane ustrukturyzowane): na przykład
 - 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 (na przykład hasła encyklopedyczne w poszczególnych wersjach językowych Wikipedii): na przykład
 - 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 (na przykład 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 skład której 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, na przykład populacji Warszawy w odniesieniu do danego roku z podaniem źródła przejęcia tych danych. Taka sytuacja ma miejsce przykładowo dla rekordu na temat Berlina (Q64)²⁰, gdzie informacje na temat populacji sięgają XIII wieku.



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

Wartości dla atrybutów są formalnie specyfikowane, to znaczy model danych zakłada, że może to być między innymi ciąg znaków, współrzędne geograficzne, jednostki

¹⁹ F. E r x l e b e n, M. G ü n t h e r, M. K r ö t z s c h, J. M e n d e z, D. V r a n d e ĩ ć, dz. cyt., s. 53.

²⁰ Zob. <https://www.wikidata.org/wiki/Q64>.

miar i wag czy też adresy URL obiektów zarówno wewnątrz bazy, jak i do zewnętrznych źródeł informacji (na przykład 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 różnych 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 między innymi 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 uszczegółowiającego 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 języku. 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 między innymi kartotek haseł wzorcowych jest realizowana za pomocą szablonu o dosyć niefortunnej nazwie *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 wzorcowych lub rekordów bibliograficznych w katalogach bibliotecznych. Celem wprowadze-

nia 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 strony stanowiło 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 bibliecznych²¹.

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 bibliecznych. Są to między innymi:

- VIAF – Wirtualna Międzynarodowa Kartoteka Haseł Wzorcowych,
- LCCN – Numer Kontrolny Biblioteki Kongresu,
- ISNI – International Standard Name Identifier,
- 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 się identyfikator występującego tam hasła. Mechanizm zaimplementowany w module wyświetla te dane jako hiperłącza do bezpośredniej lokalizacji hasła w danej kartotece. Na przykład zawartość modułu dla hasła w Wikipedii *John Flanagan (pisarz)*²² zawiera między innymi:

```
{{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:



Rysunek 2. Moduł kontroli autorytatywnej w Wikipedii.

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

²² [https://pl.wikipedia.org/wiki/John_Flanagan_\(pisarz\)](https://pl.wikipedia.org/wiki/John_Flanagan_(pisarz)).

W 2012 roku z inicjatywy wikipedystów zatrudnionych w OCLC i British Library (Wikipedian in Residence) uruchomiono projekt automatycznej identyfikacji haseł osobowych w Wikipedii, w bazie VIAF. W efekcie powstało oprogramowanie VIAFbot²³, które na podstawie zastosowanych algorytmów mapowania umieściło identyfikatory VIAF dla ponad 250 tysięcy 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 około 10–15% przypadków. Uruchomiono również osobną stronę w Wikipedii, na której użytkownicy mogą zgłaszać dostrzeżone nieprawidłowości²⁴.

Dzięki integracji Wikipedii z bazą wiedzy Wikidane cały proces wypełniania treścią modułu *Kontrola autorytatywna* jest zautomatyzowany. Z punktu widzenia wikipedysty sprowadza się on wyłącznie do uruchomienia modułu²⁵ 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, między innymi w kartotekach haseł wzorcowych, katalogach bibliotecznych, dziedzinowych bazach danych oraz dokumentach normalizacyjnych. W ramach tej klasy wyodrębniono dodatkowych 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 języku 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

²³ A. K y r i o s, *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].

²⁴ Zob. <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: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
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 hasel (148) zgrupowano w klasie Q18614948. Pozostałe kartoteki i bazy danych przyporządkowano do klas reprezentujących poszczególne kategorie hasel. Z danych przedstawionych w Tab. 1 wynika, że najwięcej źródeł informacji w celu szeroko rozumianej kontroli hasel 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 hasel wzorcowych bibliotek narodowych, na przykład:

- 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 hasel osobowych bazy 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 była identyfikacja rekordów haseł osobowych w bazie Wikidane, w których występują identyfikatory z wyżej wymienionych kartotek haseł wzorcowych. Dla każdego hasła odnotowano również jego obecność polskojęzycznej wersji Wikipedii. W celu pozyskania danych wykorzystano interfejs programistyczny Wikidanych w postaci protokołu 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 roku.

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*).

W tabeli 3 przedstawiono rozkład ilościowy haseł osobowych w bazie Wikidane zawierających identyfikatory kartotek haseł wzorcowych wykorzystane 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. Zakres wykorzystania kartotek haseł wzorcowych w hasłach osobowych w bazie Wikidane

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 Wikidane 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 unikatowych identyfikatorów. O ile baza Wikidane zapewnia ujednoczony 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 w postaci numeru systemowego rekordu (pole 001 MARC 21) oraz numeru kontrolnego rekordu (pole 010 MARC 21). Na przykład 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 nie tyle lokalizuje rekord (zasób informacyjny) w strukturze plików na serwerze, co zawiera formalnie wyrażone zapytanie o wyświetlenie rekordu wzorcowego o identyfikatorze 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 tak zwanych klastrach VIAF²⁶. Tym samym integracja NUKAT jako źródła informacji dla haseł w bazie Wikidane odbywa się za pośrednictwem VIAF. Taki sposób identyfikacji haseł wzorcowych z bibliotecznych khw w Wikidanych dotyczy również na przykład 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). Choć 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).

²⁶ Zob. M. R o s z k o w s k i, *Kartoteka haseł wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF*, [w:] *Bibliografia – teoria, praktyka, dydaktyka*, red. J. W ó ź n i a k - K a s p e r e k, J. F r a n k e, Warszawa 2016, s. 203–222.

Kolejnym etapem badań była próba udzielenia odpowiedzi na pytanie o liczbę haseł osobowych w polskojęzycznej Wikipedii, w których to hasłach występują identyfikatory haseł wzorcowych z siedmiu wspomnianych kartotek haseł wzorcowych. Dane uzyskane za pośrednictwem bazy Wikidane 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. W 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 na przykład:

- 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 miał miejsce. Można przypuszczać, że w momencie pobierania danych z VIAF w tej bazie nie istniały jeszcze rekordy przesłane przez NUKAT, stąd brak tych danych w bazie Wikidane.

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.

* * *

Integracja bibliotecznych kartotek hasel wzorcowych w semantycznych bazach danych, czy też szerzej – w projektach realizujących założenia sieci semantycznej, wymaga ich przygotowania pod względem formalnym, to znaczy spełnienia wymogów usługi sieciowej²⁷ 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 informacji, które przez lata prowadzenia przez biblioteki osiągnęły wysoki poziom jakości prezentowanych informacji w środowisku sieciowym jako pełnoprawny element tak zwanej sieci danych. Kartoteki hasel 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 zbioru słownictwa kontrolowanego, które zapewnią pożądaną poziom spójności danych.

THE INTEGRATION OF THE PERSONAL AUTHORITY RECORDS IN THE WIKIDATA SEMANTIC KNOWLEDGE BASE

SUMMARY

This article deals with the problem of the presence of personal authority records in a network environment from the point of view of the so-called data network (Web of Data). Its aim is to show – on the example of integration in the Wikidata semantic knowledge base – how these records (the so-called ‘khw’ records) function outside the context of a library catalogue. The purpose of this research has been to specify the scope of the use of library ‘khw’ records as sources of information that allow us to control the forms of personal names in the Wikidata database. For this purpose, the author has analysed the data model in the Wikidata database and identified those metadata elements in it which are responsible for the implementation of the basic functions of personal authority records. Additionally, the author discusses the use of seven ‘khw’ records in the Wikidata database – also with regard to the Polish language Wikipedia.

²⁷ Zob. np. M. R o s z k o w s k i, *Kartoteka hasel wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF*. [w:] *Bibliografia – teoria, praktyka, dydaktyka*, red. J. W o ź n i a k - K a s p e r e k, J. F r a n k e. Warszawa 2016, s. 203–222.

SŁOWA KLUCZOWE:

Bazy danych, kartoteki haseł wzorcowych, metadane, Sieć Semantyczna, Wikidane, Wikipedia

KEYWORDS:

Databases, personal authority records, metadata, Semantic Web, Wikidata, Wikipedia