

**Włodzimierz Lewoniewski, Aleksandra Kasprzak,
Krzysztof Węcel, Witold Abramowicz**

Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektro-
nicznej, Katedra Informatyki Ekonomicznej

Autor do korespondencji: Włodzimierz Lewoniewski,
włodzimierz.lewoniewski@ue.poznan.pl

**KOMPLETNOŚĆ DANYCH O PRODUKTACH
W INFOBOKSACH RÓŻNYCH WERSJI
JĘZYKOWYCH WIKIPEDII**

Streszczenie: Wikipedia jest obecnie jednym z najpopularniejszych źródeł wiedzy. Wiele artykułów tej wielojęzycznej encyklopedii zawiera informacje o różnego rodzaju produktach, np. samochodach, filmach, grach komputerowych, telefonach komórkowych. Artykuły o każdym z tych dóbr mogą powstawać niezależnie w różnych językach, w związku z czym można zaobserwować różnice w treściach opracowań poszczególnych użytkowników (w tym anonimowych). Często w artykułach umieszcza się wyróżnioną ramkę, która w przejrzysty sposób ma prezentować najważniejsze informacje o podmiocie artykułu, tzw. infoboks. Czytelnicy korzystają z takich ramek, aby uzyskać najważniejsze informacje o produkcie bez analizy treści całego artykułu. Dodatkowo dane z infoboksów mogą służyć do wzbogacania innych popularnych baz danych, takich jak np. DBpedia. Z tego powodu szczególnie istotna jest weryfikacja jakości wprowadzanych przez użytkowników danych, gdzie jakość może być charakteryzowana przez takie miary, jak aktualność, poprawność i kompletność. W niniejszym artykule zostanie przeprowadzona analiza kompletności danych o produktach z różnych grup, pochodzących z różnych wersji językowych Wikipedii. Wyniki badań mogą znaleźć zastosowanie w praktyce do wzbogacenia informacji o produktach.

Słowa kluczowe: Wikipedia, DBpedia, jakość danych, infoboks, kompletność, opis produktu.

Klasyfikacja JEL: C55, D8, L15, L86.

THE COMPLETENESS OF PRODUCT DATA IN INFOBOXES IN DIFFERENT LANGUAGE VERSIONS OF WIKIPEDIA

Abstract: Wikipedia is currently one of the most popular sources of knowledge. Many articles in this multilingual encyclopaedia contain information about different products, such as cars, movies, computer games or mobile phones. Articles about these goods can be edited independently in various languages, and differences in the content added by users (including anonymous ones) can be observed. An article usually contains an infobox, which in a concise way presents the most important facts from the article. Readers can use infoboxes to get the most important information about a product without analysing the content of the entire article. Additionally, data extracted from an infobox can be used to enrich other popular databases, such as DBpedia. For this reason, it is particularly important to verify the data quality provided by users, where quality can be characterized by metrics such as timeliness, correctness and completeness. This paper examines the completeness of product data from various groups from different language versions of Wikipedia. The results of the study can be applied in practice to enrich information about products.

Keywords: Wikipedia, DBpedia, data quality, infobox, completeness, product.

Wstęp

Przez 15 lat od czasu powstania Wikipedia zdobyła pozycję jednego z ważniejszych źródeł ogólnodostępnej informacji encyklopedycznej. Jej cechą charakterystyczną jest to, że jest współtworzona przez wielu użytkowników. Obecnie Wikipedia jest na piątym miejscu w rankingu najczęściej odwiedzanych stron w Internecie (Alexa, 2018), ustępując tylko Google, YouTube, Facebookowi oraz Baidu. Wikipedia zawiera ponad 44 mln artykułów w ponad 290 różnych językach (Meta-Wiki, 2018). Największa jest angielska wersja językowa (EN), która zawiera ponad 5,4 mln artykułów. Do najbardziej rozwiniętych wersji językowych należą również niemiecka (DE) z ponad 2 mln artykułów, a także francuska (FR), rosyjska (RU) i polska (PL) z ponad 1 mln artykułów każda.

Zmiany wprowadzane przez użytkowników do każdego artykułu są zapisywane w historii edycji, która pozwala na śledzenie zmian i umożliwia przywrócenie zawartości artykułu do poprzedniej wersji. Do kwietnia 2017 r. użytkownicy dokonali łącznie ponad 2,3 mld edycji we wszystkich wersjach językowych (Meta-Wiki, 2018).

W odróżnieniu od innych popularnych serwisów internetowych Wikipedia nie wyświetla reklam i utrzymuje się z darowizn od użytkowników. Według niektórych szacunków Wikipedia mogłaby zarobić na reklamie ponad 2,3 mld USD rocznie (Johnston, 2013). W 2016 r. koszty utrzymania tej encyklopedii wyniosły około 66 mln USD, podczas gdy przychód z darowizn wyniósł ponad 77 mln USD (Wikimedia Foundation, 2016).

Pomimo niekomercyjnego charakteru Wikipedii informacje zawarte w tej bazie wiedzy tworzonej przez społeczność mogą wpływać na decyzje biznesowe i konsumenckie. Strony Wikipedii o znanych osobach, firmach czy produktach często pojawiają się jako pierwsze w wynikach wyszukiwania Google, Bing, Yandex i innych popularnych serwisów. Można się spodziewać, że osoby odwiedzające strony Wikipedii oraz jej twórcy są zainteresowani wysoką jakością treści w niej zawartej. Firmy monitorują strony Wikipedii dotyczące ich samych lub ich produktów, aby mieć pewność, że nie pojawiają się tam informacje niepożądane. Dość ciekawa sytuacja zdarzyła się w 2017 r., kiedy to sieć restauracji Burger King postanowiła zareklamować swój produkt, używając słów kluczowych, które automatycznie uruchamiały aplikacje wyszukiwawcze na urządzeniach z zainstalowanym oprogramowaniem Google Assistant. Reklama wymuszała na urządzeniach wyszukiwanie informacji o produkcie firmy Burger King, m.in. w Wikipedii. W tym czasie opis tego produktu w Wikipedii został zmieniony przez jednego z użytkowników – we wprowadzeniu było zaznaczone, że jest to „najgorszy produkt”, a także dodane inne niesprawdzone informacje (Wakefield, 2017).

Artykuły o określonych produktach mogą powstawać niezależnie w dowolnej wersji językowej Wikipedii. W związku z tym jakość informacji o tym samym produkcie może się różnić w zależności od języka. Opis produktu w jednej wersji językowej nie zostanie przetłumaczony na inne języki bez inicjatywy użytkowników Wikipedii.

Możemy się spodziewać, że największa – anglojęzyczna – wersja Wikipedii posiada najwięcej informacji o produktach. Jednak przedstawienie informacji o produkcie w różnych językach jest szczególnie ważne dla konsumentów, którzy korzystają z wyszukiwarek w swoim ojczystym języku, niekoniecznie angielskim. Poza tym niektóre produkty mogą być bardziej popularne na rynkach lokalnych, a zatem większe będzie prawdopodobieństwo znalezienia informacji na ich temat w odpowiednich wersjach językowych. Opisy produktów w lokalnych językach mogą też być interesujące dla firm, które chcą je odpowiednio pozycjonować na rynku.

Niektóre artykuły Wikipedii mają dołączoną specjalną wyróżnioną ramkę, umieszczaną zazwyczaj w prawym górnym rogu artykułu, która w przejrzysty sposób prezentuje najważniejsze informacje. Jest to infoboks. Dane z infoboksów mogą służyć nie tylko do szybkiego zapoznania się z tematem przez czytelnika Wikipedii, ale również są wygodniejsze do przetwarzania maszynowego. Mogą zatem służyć do wzbogacenia innych popularnych baz danych, takich jak np. DBpedia (DBpedia, b.d.; Lehmann i in., 2015). Z tego powodu szczególnie istotna jest weryfikacja jakości wprowadzanych przez użytkowników treści, gdzie jakość może być charakteryzowana przez takie miary, jak aktualność, poprawność i kompletność.

1. Jakość Wikipedii

Wikipedia jest często krytykowana za niską jakość zawartych w niej informacji (Krytyka Wikipedii, b.d.). Jest to głównie pochodną zasady wolności tworzenia i edycji artykułów. W odróżnieniu od tradycyjnych encyklopedii w Wikipedii artykuły mogą współtworzyć nawet niezarejestrowani anonimowi użytkownicy, którzy nie muszą formalnie potwierdzać swoich kompetencji i doświadczeń w określonych dziedzinach.

W ujęciu ekonomicznym informację można rozpatrywać jako produkt, który musi spełniać określone kryteria jakości. W zależności od wiedzy, doświadczenia i popytu informacyjnego każdy użytkownik może inaczej oceniać jakość tego produktu (Lewoniewski, Węcel i Abramowicz, 2015). Przy ocenianiu subiektywnym szczególnie ważne jest precyzyjne definiowanie tych atrybutów, aby oceniający jasno rozumiał, co podlega ocenie (Grudzień, 2012). W związku z tym dla dokonania oceny jakości informacji należy brać pod uwagę jakość wszystkich jej atrybutów (Abramowicz, 2008).

W odpowiedzi na krytykę społeczność Wikipedii utworzyła system wyróżniania artykułów o wysokiej jakości. Praktycznie każda wersja językowa tej encyklopedii posiada specjalne oznakowanie dla najlepszych artykułów: w polskojęzycznej Wikipedii oznacza się je jako „Artykuły na medal” (ANM), w angielskiej – „Featured Article” (FA). Takie artykuły muszą spełniać określone standardy: wyczerpujący opis tematu wraz z ilustracjami, właściwy przegląd źródeł, właściwy styl pisania (z zasadami neutralnego punktu widzenia, weryfikowalności i praw autorskich) i inne. Jeżeli artykuł spełnia najważniejsze kryteria artykułu wysokiej jakości, jednak nie jest do końca zgodny z zasadami ANM, może otrzymać nieco niższe wyróżnienie – „Dobry Artykuł” (DA). W angielskiej wersji jest to „Good Article”

(GA). W tak ocenionych artykułach możemy się spodziewać wystarczającego opisu najważniejszych zagadnień z wykorzystaniem istotnych źródeł (Porównanie wyróżnień artykułów, b.d.). Należy zaznaczyć, że najlepszych artykułów, tj. odpowiedników angielskiego wyróżnienia FA i GA, jest bardzo mało – zazwyczaj około 0,5% wszystkich artykułów w konkretnej wersji językowej.

Istnieją również inne oceny jakości, które może otrzymać artykuł. Wskazują one najczęściej na stopień „dojrzałości” artykułu. Na przykład w polskiej Wikipedii są to: Czwórka, Poprawny, Dostateczny, Start, Załączek. Niestety ponad 95% artykułów polskiej Wikipedii nie posiada żadnej oceny jakości.

Istnieje szereg badań zajmujących się metodami automatycznej oceny jakości artykułów Wikipedii. Taką jakość można oceniać na podstawie treści artykułu (Warncke-Wang, Cosley i Riedl, 2013; Węcel i Lewoniewski, 2015; Lewoniewski, Węcel i Abramowicz, 2016), historii edycji (Ingawale, Dutta, Roy i Seetharaman, 2013), strony dyskusji tego artykułu oraz innych źródeł. Serwis internetowy WikiRank używa niektórych miar ilościowych (jak długość tekstu, liczba obrazków, referencji) w celu oceny względnej jakości artykułów z Wikipedii w różnych językach (WikiRank, b.d.).

Wysoka jakość artykułu nie oznacza jednak wysokiej jakości jednych z najważniejszych jego elementów – infoboksów (Węcel i Lewoniewski, 2015). W literaturze stosunkowo mało uwagi poświęca się analizie jakości tej ustrukturyzowanej części artykułów w różnych wersjach językowych Wikipedii. Jakość infoboksu można rozpatrywać z uwzględnieniem miar jakości artykułu, w którym został on umieszczony (Lewoniewski, 2017). Inne badania oceniają jakość danych pochodzących z infoboksów w kontekście baz wiedzy, które są wzbogacane automatycznie na podstawie danych z tych infoboksów – np. DBpedia. Istnieją różne metody i narzędzia do oceny jakości w tej semantycznej bazie danych. Jedno z takich rozwiązań – RDFUnit, które wykorzystuje predefiniowane wzorce testów jakości oparte na szablonie zapytań SPARQL do analizy integralności danych (Kontokostas i in., 2014). Framework Luzzu jest oparty na ontologii i umożliwia wdrażanie różnych mar bez konieczności stosowania zapytań SPARQL (Debattista, Auer i Lange, 2016). Ponieważ DBpedia jest przedstawicielem Linked Open Data (LOD), do łączenia danych z różnych języków można zastosować rozwiązanie zaproponowane przez Sieve (Mendes, Mühleisen, i Bizer, 2012). Istnieją również algorytmy, które mogą identyfikować brakujące typy stwierdzeń oraz błędne twierdzenia w LOD (Paulheim i Bizer, 2014). Niemniej jednak metody te wymagają zaangażowania ekspertów

dziedzinowych w celu identyfikacji miar oceny jakości, a dla bardziej dogłębnej analizy jakości LOD konieczne jest uwzględnienie dodatkowych wymiarów jakości (Zaveri i in., 2016).

Problemem badawczym przedstawianym w niniejszym artykule jest analiza jednego z ważniejszych aspektów jakości infoboksów – kompletności. Analiza ze względu na ekonomiczny charakter artykułu została zawężona do infoboksów dotyczących produktów. Z wielu dostępnych kategorii produktów wybrano te, które były najpopularniejsze, a tym samym mogły dostarczyć odpowiedni materiał do badań. Analiza została przeprowadzona dla sześciu kategorii: albumy muzyczne, gry komputerowe, telefony komórkowe, samochody, filmy i oprogramowanie. Przeanalizowano artykuły w siedmiu wersjach językowych: angielskiej (EN), niemieckiej (DE), francuskiej (FR), polskiej (PL), rosyjskiej (RU), ukraińskiej (UK) i białoruskiej (BE). Taki wybór języków wynikał z kompetencji językowych autorów. Dodatkowo w pierwszych pięciu językach napisano stosunkowo dużo artykułów w Wikipedii – każdorazowo był to co najmniej milion. Pozostałe, uboższe wersje językowe służyły jako punkt wyjścia do wzbogacania kompletności artykułów.

Jako główną metodę badawczą zastosowano analizę ilościową, w tym w szczególności uczenie maszynowe oraz wizualizację. Analiza została przeprowadzona na kompletnych tekstach artykułów Wikipedii, bez historii edycji, dostępnych w postaci archiwów na stronie Wikimedia (Wikimedia Downloads, 2018).

W dalszych rozdziałach przybliżane jest pojęcie infoboksu oraz opisany jest problem łączenia wartości atrybutów w różnych wersjach językowych. W kolejnej sekcji przedstawiona została charakterystyka produktów w Wikipedii. W sekcji 5 zaproponowane jest rozwiązanie problemu badawczego. Artykuł kończy podsumowanie wyników naukowych oraz przedstawienie implikacji dla praktyki gospodarczej.

2. Infoboksy

Od strony technicznej infoboks to szablon, który jest definiowany przez użytkowników Wikipedii, a następnie, odpowiednio wypełniony, umieszczany w artykule. Zmiana w kodzie szablonu automatycznie pociąga za sobą zmiany we wszystkich artykułach z niego korzystających – to pozwala zachować spójny wygląd infoboksów danego typu.

Szablon infoboksu składa się z elementów dwóch rodzajów: atrybutów i wartości atrybutów. Infoboks określonego typu ma ściśle określony

zestaw atrybutów, których można używać do opisu określonego podmiotu czy wydarzenia. Na przykład w infoboksie opisującym samochody można wykorzystać atrybut „zbiornik paliwa”, natomiast w infoboksie o telefonach komórkowych ten atrybut zostanie zignorowany i nie będzie wyświetlony. Na rysunku 1 przedstawiony został przykład wypełnionego infoboksu o samochodzie.

The diagram illustrates the process of rendering a Wikipedia infobox. On the left, a raw code snippet is shown, containing various parameters for a car. On the right, the rendered infobox is displayed, showing the car's name, image, and various specifications. A large arrow points from the code to the rendered version. A label 'Parametr' points to the 'Zaprezentowany' field in the rendered infobox.

Raw Code (Left):

```

{{Auto infobox
|nazwa          = Foleki Fiat 125p<br />FSO 125p
|zdjęcie        = Foleki_Fiat_125p_1980.jpg
|opis zdjęcia   = Foleki Fiat 125p MR'80 zarejestrowany na
|[[tablica rejestracyjna|tablicach pojazdu zabytkowego]]
|inne nazwy     = FSO 1300/1500<br />FSO Folda<br />FSO
Montana<br />FSO Arizona<br />FSO Classic<br />FSO Linda<br />Nasr
125<br />Zastava 125p<br />"duży Fiat" <small>
|[[kolokwializm|kolokwialnie]]</small>
|producent      = [[Fabryka Samochodów Osobowych|FSO]]<br
/>[[Nasr]]<br />[[Zastava]]<br />CCA (Kolumbia)
|projektant     =
|premiera       = 1967
|okres produkcji = 28 listopada 1967 – 29 czerwiec 1991
|miejsce produkcji = [[flaga|POL]] [[Warszawa]]<br />montaż:<br
[[flaga|EGY]] [[Kair]]<br />[[flaga|YUG]]warzant<sup>3</sup>FR]] [[Kragujevac]]<br
/>[[flaga|COL]] [[Bogota]]<br />[[flaga|IDN]] [[Dakarta]]<br
/>[[flaga|THA]] [[Bangkok]]
|poprzednik     = [[FSO Warszawa|FSO Warszawa 223/224]]
|następca       = [[FSO Polonez|FSO Polonez MR'78]]
|segment        = [[Autosegment D|D]]
|typy nadwozia  = 4-drzwiowy [[sedan (nadwozie)|sedan]]<br
/>5-drzwiowe [[kombi (nadwozie)|kombi]]<br />2-drzwiowy [[pick-up]]
|silniki         = podstawowe: [[OHV]] 1300, 1500<br
/>pozostałe: [[OHV]] 1600
|skrzynia biegów = 4-biegowa manualna<br />5-biegowa manualna
|rodzaj napędu   = [[napęd tylny|tylny]]
|długość        = 4226 mm (sedan) – starsza wersja z klami 4232
mm <br />4234 mm (kombi)
|szerokość      = 1630 mm
|wysokość       = 1440 mm
|rozstaw osi    = 2506 mm
|masa własna     = od 970 kg
|zbiornik paliwa = 45 l
|liczba miejsc   = 5<br />2 (pick-up)
|gwiazdki EuroNCAP =
|pojemność bagażnika = 360 l (sedan)<br />1100 / 1550 l (kombi)
}}

```

Rendered Infobox (Right):

Polski Fiat 125p
FSO 125p

Polski Fiat 125p MR80 zarejestrowany na tablicach pojazdu zabytkowego

Inne nazwy	FSO 1300/1500 FSO Folda FSO Montana FSO Arizona FSO Classic FSO Linda Nasr 125 Zastava 125p duży Fiat (kolokwialnie)
Producent	FSO Nasr Zastava CCA (Kolumbia)
Zaprezentowany	1967
Okres produkcji	28 listopada 1967 – 29 czerwiec 1991
Miejsce produkcji	Warszawa montaż: Kair Kragujevac

Rysunek 1. Infoboks opisujący samochód (z lewej strony – kod źródłowy dla osoby edytującej artykuł, z prawej – wersja dla czytelników Wikipedii)

Wstawianie infoboksu do artykułu nie jest obowiązkowe, ale jest zalecane dla stron określonego typu, dla których infoboksy zostały przewidziane. Zatem możemy się spodziewać, że w artykułach wysokiej jakości (ANM i DA) infoboksy często będą umieszczone.

2.1 Kompletność infoboksu

Atrybuty infoboksów i ich wartości zazwyczaj są wprowadzane przez użytkowników, którzy mają różne doświadczenie i wiedzę na określony temat. Zatem wymagane jest dodatkowe sprawdzenie jakości tych danych przez bardziej doświadczonych redaktorów.

Jakość danych może zależeć od różnych czynników, które można pogrupować w wymiary: kompletność, aktualność, dokładność i inne (Heinrich i Klier, 2015). Kompletność K infoboksu można mierzyć jako stosunek liczby wypełnionych wartościami atrybutów do liczby wszystkich atrybutów w infoboksie danego typu (Zaveri i in., 2016):

$$K = \frac{WA}{A},$$

gdzie:

WA – liczba wypełnionych atrybutów,

A – liczba wszystkich atrybutów.

2.2. Unifikacja nazw atrybutów

Infoboksy określonego typu często mają swoje odpowiedniki w różnych językach. Na przykład infoboks opisujący samochody w polskiej wersji ma nazwę „Auto infobox”. Jego odpowiednik w niemieckiej to „Infobox PKW-Modell”, a w angielskiej – „Infobox automobile”. Różnice również można zauważyć w zbiorze oraz nazwach atrybutów. Może to utrudniać porównanie wartości atrybutów infoboksu między różnymi wersjami językowymi tego samego produktu.

Pomocna w rozwiązaniu tego problemu może być semantyczna baza wiedzy – DBpedia, która odwzorowuje atrybuty infoboksów na specjalną ontologię i tym samym umożliwia wskazywanie ekwiwalentnych atrybutów w różnych językach (Bizer i in., 2009). Warunkiem wstępnym jest poprawny opis reguł mapowania każdego atrybutu w każdej wersji językowej (DBpedia Mappings, b.d.). Na przykład dla samochodów atrybuty „producer” w polskiej wersji, „Marke” w niemieckiej, „manufacturer” w angielskiej będą mapowane do wspólnego „manufacturer” w ontologii DBpedii.

3. Produkty w Wikipedii

W Wikipedii obok informacji na temat znanych osób, miast czy wydarzeń można również znaleźć treści związane z produktami, np. filmy, samochody, telefony komórkowe. Każdy produkt może być opisany w różnych językach. W niniejszej pracy rozpatrujemy sześć grup produktów w siedmiu wersjach językowych Wikipedii.

Mechanizm wyszukiwania w Wikipedii pozwala znaleźć wszystkie artykuły, które zawierają określony infoboks. W celu znalezienia artykułów na określony temat zidentyfikowane zostały nazwy infoboksów przypisane do jednej z sześciu analizowanych kategorii produktów w różnych językach. Zidentyfikowane artykuły w każdym języku dodatkowo były analizowane pod kątem posiadania innych wersji językowych, dzięki czemu można było uwzględnić artykuły o produktach, które nie posiadały szukanych infoboksów.

Warto zaznaczyć, że wybór infoboksów jako kryterium wyszukiwanego artykułów był celowy. W Wikipedii artykuły są również klasyfikowane według systemu kategorii, jednak przypisanie kategorii wymaga od edytora umieszczenia w treści artykułu dodatkowego kodu. Artykuły mogą być przypisywane do szerokiego zakresu kategorii, nie zawsze tematycznie powiązanych z ich treścią. Dodatkowo kategorie są budowane w taki sposób, że przechodząc do podkategorii, można przejść do artykułów innego rodzaju: np. w kategorii „Gry komputerowe” można razem z podkategoriami „Gry komputerowe według platformy” oraz „Gry komputerowe według roku wydania” znaleźć m.in. „Wydawcy gier komputerowych”, „Producenci gier komputerowych” czy też „Prawo i cenzura gier komputerowych”. Kategorie nie tworzą zatem taksonomii, a więc nie można wnioskować o typach artykułów.

W celu ekstrakcji nazw artykułów o konkretnych produktach w badanych wersjach językowych Wikipedii oraz ekstrakcji i analizy atrybutów infoboksów zostały opracowane specjalne narzędzia w języku Python (niektóre z tych narzędzi są dostępne na stronie GitHub, b.d.).

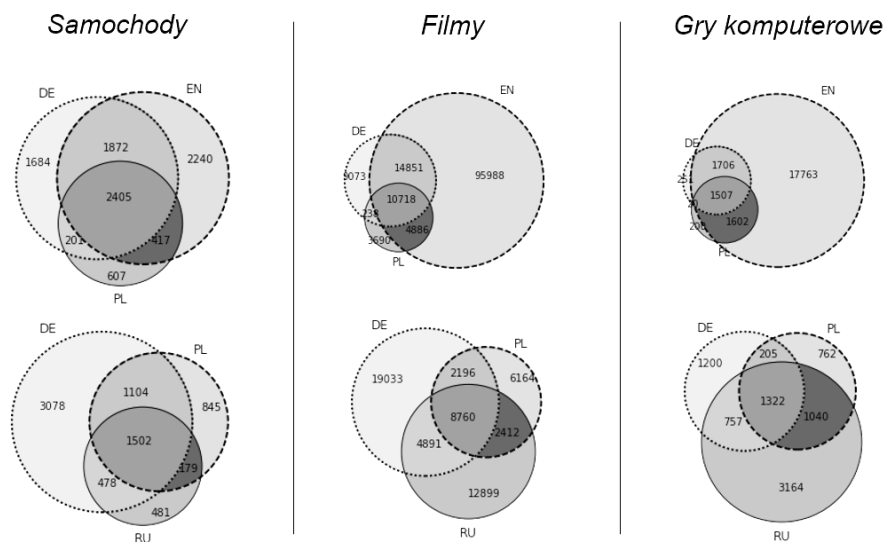
Tabela 1. Artykuły o produktach w Wikipedii (dane na czerwiec 2018 r.)

Temat	BE	DE	EN	FR	PL	RU	UK	Unik.
Albumy muzyczne	370	14 107	156 902	43 228	29 353	20 991	10 104	181 015
Samochody	220	6 119	6 874	4 621	3 598	2 615	2 335	25 805
Filmy	300	34 813	126 332	51 700	19 489	28 917	15 605	1 944
Telefony komórkowe	11	271	1 490	340	610	619	492	10 600
Oprogramowanie	109	4 682	11 663	4 534	2 498	4 743	2 244	156 825
Gry komputerowe	122	3 459	22 531	12 912	3 290	6 246	1 712	14 221

Źródło: Na podstawie statystyk Wikipedii.

Liczby artykułów o produktach każdego typu w każdej wersji językowej są przedstawione w tabeli 1. W ostatniej kolumnie została podana liczba unikatowych produktów, które są opisane w co najmniej jednej wersji językowej Wikipedii.

Można zauważyć, że bardziej rozwinięte wersje językowe zazwyczaj posiadają więcej artykułów o produktach. Z punktu widzenia weryfikacji i wzbogacania danych szczególnie interesujące są artykuły napisane w więcej niż jednym języku. Pokrycie w różnych językach zostało przeanalizowane za pomocą diagramów Venna. Rysunek 2 przedstawia analizę wybranych kategorii produktów według trójek języków EN-DE-PL oraz DE-PL-RU (w celu oszczędzania miejsca zostały przedstawione tylko niektóre tematy oraz wersje językowe. Bardziej szczegółowa analiza jest przedstawiona na stronie Lewoniewski.info, b.d.).



Rysunek 2. Pokrycie artykułów o produktach w wersjach językowych Wikipedii (dane na czerwiec 2018 r.)

Otrzymane diagramy pokazują, że często najwięcej artykułów mieści się w największej, angielskiej wersji językowej Wikipedii. Pewnym wyjątkiem są artykuły o samochodach, gdzie wersje niemiecka i angielska mają zbliżoną liczebność. Dodatkowo sprawdzono artykuły pod kątem posiadania infoboksu. W niektórych wersjach językowych ponad 30% artykułów określonej grupy produktów nie posiadało infoboksów (zob. tabela 2).

Tabela 2. Udział artykułów bez infoboksów

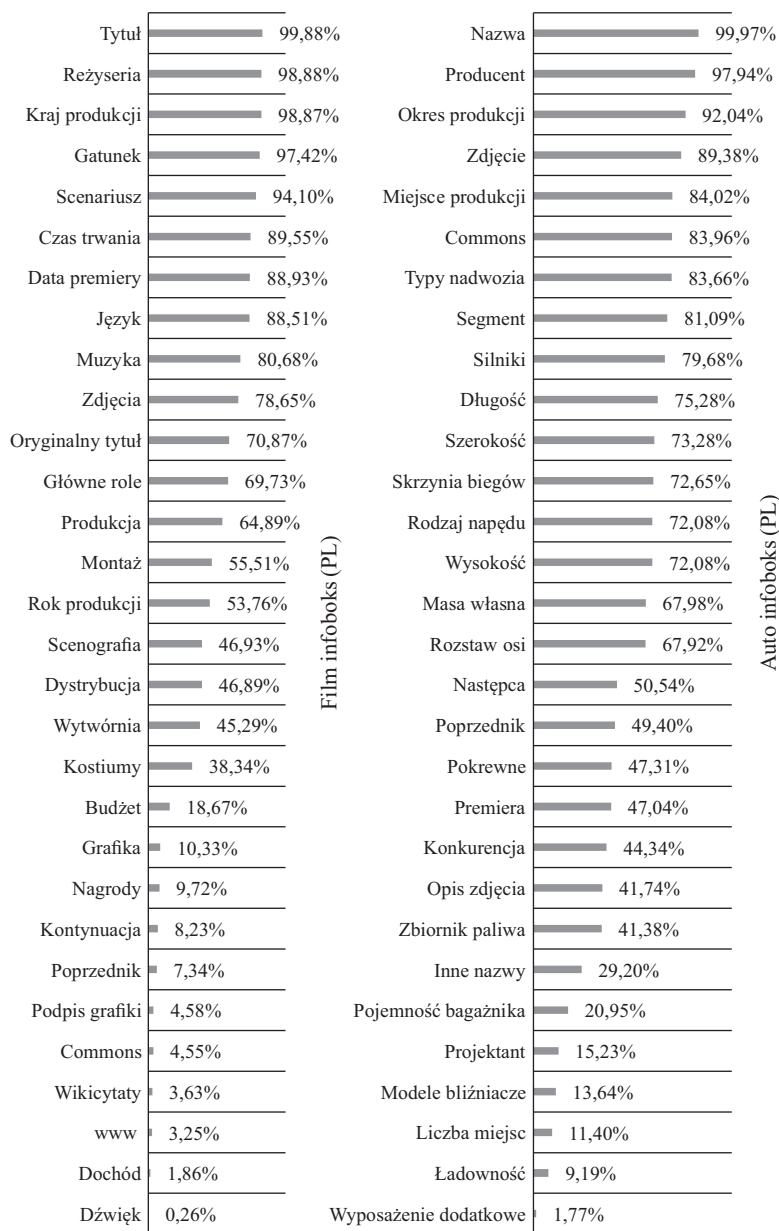
Temat	BE	DE	EN	FR	PL	RU	UK
	%						
Albumy muzyczne	49,73	36,95	7,24	8,65	21,43	29,08	27,22
Samochody	34,55	37,88	9,25	8,81	7,12	7,19	96,32
Filmy	16,33	0,92	3,26	21,95	1,19	6,52	2,36
Telefony komórkowe	27,27	23,99	7,38	27,94	25,41	14,22	6,50
Oprogramowanie	38,53	17,71	7,59	26,31	37,31	16,85	23,13
Gry komputerowe	54,92	10,61	3,09	2,81	6,44	4,58	9,05

Źródło: Obliczenia własne w czerwcu 2018 r. na podstawie (Wikimedia Downloads, 2018).

Należy dodać, że niektóre artykuły nieposiadające bezpośrednio tematycznego infoboksu mogły mieć inny podobny szablon. Na przykład niektóre artykuły o grach zawierały infoboks „Seria gier komputerowych”, który opisuje ogólnie kilka gier bez podania szczegółowej informacji o każdej (nazwa, data wydania, platforma, wydawca itp.). Zidentyfikowano również sytuacje, kiedy w jednej wersji językowej była opisana gra komputerowa, a po przejściu do innej wersji językowej opis dotyczył tematycznie powiązanego filmu, w którego opisie było wspomniane istnienie gry o podobnej nazwie. Stosunkowo rzadko można spotkać artykuły, które posiadają dwa lub więcej infoboksów. Na przykład „Star Trek” jest i grą komputerową, i filmem. Zatem może posiadać on dwa infoboksy, które będą umieszczone w różnych miejscach. W naszych badaniach braliśmy pod uwagę tylko pierwszy (górny) infoboks.

4. Kompletność danych o produktach

Kompletność infoboksu została już zdefiniowana jako stosunek liczby wypełnionych atrybutów do liczby wszystkich atrybutów. W niniejszym rozdziale przedstawione są wyniki analiz, w jakim stopniu autorzy Wikipedii wypełniają poszczególne atrybuty w infoboksach. Jeśli chodzi o polskojęzyczną Wikipedię i kategorie produktów samochody oraz filmy, to prawie w 100% przypadków w takich infoboksach jest uzupełniany atrybut „nazwa” („tytuł”). Bardziej specyficzne atrybuty uzupełniane są znacznie



Rysunek 3. Częstość wypełniania atrybutów w infoboksach o samochodach i filmach w polskojęzycznej Wikipedii (dane na czerwiec 2018 r.)

rzadziej, np. „Ładowność”, „Projektant”, „Dochód” mogą posiadać wartość w mniej niż 20% wszystkich wystąpień danego infoboksu. Szczegółowe wyniki częstości wypełniania poszczególnych atrybutów dla dwóch kategorii produktów w polskiej Wikipedii przedstawiono na rysunku 3.

W kolejnej analizie porównano listy atrybutów odpowiadających sobie infoboksów w różnych językach. Najpierw porównano liczebności różnych atrybutów w infoboksach tego samego typu. Wyniki przedstawione w tabeli 3 pokazują różną „kulturę” w definiowaniu dostępnych atrybutów w infoboksach danego typu w siedmiu badanych wersjach językowych Wikipedii.

Tabela 3. Liczba predefiniowanych atrybutów w infoboksach (brano pod uwagę atrybuty, które posiadały wartości w co najmniej 5 infoboksach danego typu)

Temat	BE	DE	EN	FR	PL	RU	UK
Albumy muzyczne	26	69	59	52	68	78	72
Samochody	28	24	71	55	32	50	39
Filmy	22	64	176	20	33	83	142
Telefony komórkowe	10	51	91	36	19	106	84
Oprogramowanie	24	33	48	26	16	68	74
Gry komputerowe	50	42	37	29	33	145	110

W tabeli 4 przedstawiono po 13 najczęściej wypełnianych atrybutów w 4 grupach produktów w każdym języku.

Analizując infoboksy pod kątem częstotliwości wypełniania atrybutów w różnych językach, również możemy dostrzec różnice. Na przykład każda wersja językowa może mieć swoją preferowaną listę atrybutów, które są wypełniane stosunkowo częściej, niż się to dzieje w innych językach. Interesująca jest wersja rosyjska, gdzie ten sam atrybut „Tytuł” może mieć dwie nazwy (zapisane alfabetem łacińskim i cyrylicą) w ramach jednego infoboksu. Dodatkowo istnieją atrybuty, które z różnych powodów nie są w ogóle wykorzystywane w określonej wersji językowej, np. przy opisie gier komputerowych w angielskiej wersji Wikipedii swego czasu zrezygnowano z atrybutu „Version” czy też „Media” w infoboksach, choć inne edycje nadal go używają.

Z powyższego wynika, że wiele atrybutów w infoboksie nie jest wypełnianych przez użytkowników Wikipedii. Stwarza to potencjał do uzupełniania brakujących atrybutów i ich wartości poprzez przeniesienie z innych wersji językowych, o ile ich odpowiedniki zostały w nich wypełnione. W ramach niniejszej pracy przeanalizowano potencjał przeniesienia tych

Tabela 4. Najczęściej wypełniane atrybuty w infoboksach w różnych wersjach językowych Wikipedii (dane na czerwiec 2018 r.)

	DE	EN	PL	RU
Samochody	von (99,95%) Versionen (99,92%) Marke (99,87%) Klasse (99,74%) Motoren (99,71%) Modell (98,68%) Radstand (93,08%) Länge (90,72%) Gewicht (90,69%) Breite (88,48%) Bild (85,3%) bis (80,51%) Höhe (80,14%) ...	name (97,97%) manufacturer (95,6%) image (89,78%) production (88,8%) class (74,75%) engine (71,56%) body_style (69,79%) layout (61,78%) length (57,72%) transmission (56,4%) width (54,64%) wheelbase (52,45%) height (49,1%) ...	nazwa (99,97%) producent (97,94%) okres produkcji (92,04%) zdjęcie (89,38%) miejsce produkcji (84,02%) commons (83,96%) typy nadwozia (83,66%) segment (81,09%) silniki (79,68%) długość (75,28%) szerokość (73,28%) skrzynia biegów (72,65%) rodzaj napędu (72,08%) ...	название (99,96%) производитель (99,59%) годы производства (95,46%) фото (95,3%) викисклад (79,91%) тип кузова (77,52%) компоновка (73,76%) класс (71,16%) колёсная формула (71,08%) двигатель (68,32%) длина (62,05%) ширина (60,27%) высота (58,37%) ...
Albumy muzyczne	Typ (99,76%) Jahr (99,32%) Künstler (98,9%) Label (98,03%) Genre (97,47%) Titel (94,62%) Laufzeit (89,62%) Produzent (86,37%) Formate (85,32%) Nächstes (82,67%) Vorheriges (82,07%) AnzahlTitel (76,26%) Besetzung (69,84%) ...	Type (54,59%) Name (54,52%) Artist (54,32%) Released (53,99%) Genre (52,27%) Label (52,17%) This album (48,79%) type (45,4%) name (45,25%) artist (45,21%) released (45,09%) Cover (44,98%) label (43,34%) ...	rodzaj albumu (100,0%) tytuł (99,75%) wydany (99,49%) wykonawca (99,11%) gatunek (97,35%) wytwórnia (94,8%) rok wydania (89,31%) rok poprzedniego albumu (71,95%) poprzedni album (71,85%) długość (71,72%) rok następnego albumu (71,16%) następny album (71,01%) producent (66,88%) ...	Тип (99,68%) Название (99,61%) Выпущен (99,48%) Исполнитель (97,98%) Год (95,56%) Обложка (92,49%) Длительность (87,45%) Следующий (80,39%) Лебл (79,63%) Предыдущий (79,16%) След_год (78,77%) Пред_год (77,58%) Записан (70,28%) ...
Gry komputerowe	Plattform (98,96%) Genre (97,25%) Release (96,21%) Entwickler (95,89%) Spielmodi (91,0%) Titel (81,23%) Sprache (80,81%) Bedienung (79,06%) Medien (76,73%) Verleger (57,7%) PEGI (43,07%) USK (42,94%) Bild (38,32%) ...	platforms (97,37%) developer (96,34%) genre (96,21%) released (95,1%) publisher (92,71%) title (89,1%) modes (87,86%) image (84,55%) caption (43,52%) composer (33,67%) designer (31,02%) series (26,58%) producer (17,14%) ...	tytuł (99,93%) data wydania (98,08%) platforma (97,72%) producent (97,69%) gatunek (97,53%) tryby gry (94,05%) wydawca (93,89%) nośniki (66,42%) kontrolery (55,43%) kategorie wiekowe (53,02%) wymagania (40,73%) seria gier (38,62%) dystrybutor (38,23%) ...	разработчик (65,36%) жанр (59,47%) изображение (59,43%) заголовок (58,35%) издатель (54,94%) управление (51,83%) платформы (47,83%) дата выпуска (40,48%) режимы (37,26%) title (37,17%) носитель (34,44%) серия (34,3%) подпись (33,56%) ...
Filmy	FSK (54,38%) OT (53,52%) PL (53,43%) REG (53,18%) OS (52,62%) LEN (50,74%) DRB (50,55%) PRO (49,81%) DS (49,07%) KAMERA (47,81%) Originaltitel (46,41%) Erscheinungsjahr (46,41%) Produktionsland (46,35%) ...	name (99,35%) director (97,81%) released (97,31%) language (96,62%) country (94,88%) starring (89,16%) runtime (77,2%) producer (75,68%) cinematography (73,62%) writer (68,78%) music (67,71%) image (63,91%) editing (62,39%) ...	tytuł (99,88%) reżyseria (98,88%) kraj produkcji (98,87%) gatunek (97,42%) scenariusz (94,1%) czas trwania (89,55%) data premiery (88,93%) język (88,51%) muzyka (80,68%) zdjęcia (78,65%) oryginalny tytuł (70,87%) główne role (69,73%) produkcja (64,89%) ...	РусНаз (99,28%) Режиссёр (98,15%) Жанр (96,71%) Страна (96,03%) Год (95,53%) В главных ролях (93,11%) Время (91,0%) Сценарист (86,74%) Изображение (86,21%) Компания (83,59%) Оператор (79,63%) ОригНаз (79,36%) Композитор (76,19%) ...

Źródło: Obliczenia własne w czerwcu 2018 r. na podstawie (Wikimedia Downloads, 2018).

DE	EN	PL	RU
Gry komputerowe			
Platform	platforms	tytuł	разработчик
Genre	developer	data wydania	жанр
Release	genre	platforma	изображение
Entwickler	released	producer	заголовок
Spielmodi	publisher	gatunek	издатель
Titel	title	trybu gry	управление
Sprache	modes	wydawca	платформы
Bedienung	image	nośniki	дата выпуска
Medien	caption	kontrolery	режимы
Verleger	composer	kategorie wiekowe	title
PEGI	designer	wymagania	носител
USK	series	seria gier	серия
Bild	producer	dystributor	подпись
Designer	artist	język	даты выпуска
Systemminima	engine	silnik	режим
....	...	...	...

releaseDate

publisher

foaf:name

computingPlatform

developer

genre



Rysunek 4. Unifikacja atrybutów infoboksów o grach wideo w różnych wersjach językowych Wikipedii

brakujących wartości z innych języków. Poniżej przedstawiono przykład infoboksów o grach komputerowych. Rozpatrzono najczęściej wypełniane atrybuty, takie jak tytuł, data wydania, platforma, producent, gatunek, wydawca. We wszystkich czterech rozpatrywanych wersjach językowych istniała odpowiednia nazwa każdego atrybutu. Używając mapowań DBpedii, przeprowadzono unifikację nazw zgodnie z rysunkiem 4.

Posługując się zunifikowanymi nazwami, można określić, w ilu przypadkach wartości poszczególnych atrybutów mogą być przenoszone z innej wersji językowej, gdzie dany atrybut nie został wypełniony. Wyniki wzbogacenia infoboksów przedstawione zostały w tabeli 5.

Tabela 5. Udział infoboksów o grach komputerowych, w których można uzupełnić atrybuty poprzez przeniesienie wartości z innych wersji językowych Wikipedii

Atrybut unikatowy	BE	DE	EN	FR	PL	RU	UK
	%						
computingPlatform	9,1	1,1	3,3	1,0	2,3	5,9	18,3
developer	25,5	4,2	4,3	3,8	2,4	12,7	10,7
genre	10,9	2,8	4,4	2,0	2,5	19,5	13,6
releaseDate	14,5	3,8	5,5	1,9	2,0	40,3	18,2
publisher	14,5	8,7	7,9	5,3	6,2	23,9	16,4
foaf:name	18,2	18,8	11,5	0,4	0,1	5,1	20,0

Źródło: Obliczenia własne w czerwcu 2018 r. na podstawie (Wikimedia Downloads, 2018).

Dodatkowo należy zwrócić uwagę na wersję językowe artykułów, które potencjalnie mogłyby powstać w sposób automatyczny poprzez przeniesienia danych z infoboksów już istniejących wersji językowych. Potencjalna liczba nowych artykułów z infoboksami w każdym języku pokazana jest w tabeli 6.

Tabela 6. Liczba artykułów z infoboksem, które mogą powstać dzięki przeniesieniu danych z innych wersji językowych Wikipedii

Temat	BE	DE	EN	FR	PL	RU	UK
Albumy muzyczne	180 645	166 872	24 031	137 743	151 568	159 974	170 887
Samochody	10 377	4 438	3 666	5 950	6 970	7 960	8 242
Filmy	156 524	121 945	30 382	105 060	137 293	127 863	141 199

Temat	BE	DE	EN	FR	PL	RU	UK
Telefony komórkowe	1 932	1 668	448	1 598	1 330	1 321	1 447
Oprogramowanie	14 112	9 524	2 547	9 678	11 715	9 470	11 971
Gry komputerowe	25 677	22 321	3 227	12 846	22 476	19 522	24 076

Źródło: Obliczenia własne w czerwcu 2018 r. na podstawie (Wikimedia Downloads, 2018).

5. Podsumowanie: wyniki i dalsze prace

W niniejszej pracy została przedstawiona metoda analizy kompletności danych o produktach zawartych w infoboksach Wikipedii. Przeprowadzono również analizę porównawczą kompletności danych pomiędzy różnymi kategoriami produktów oraz różnymi wersjami językowymi Wikipedii. Wyniki pokazały, w jaki sposób wykorzystywanie różnych wersji językowych Wikipedii może pozwolić na automatyczne wzbogacenie danych w infoboksach w innych językach. W tym procesie istotna jest poprawna unifikacja nazw atrybutów infoboksów, które mogą mieć różne brzmienie w każdej wersji językowej.

Dodatkowo wskazano na różną „kulturę” wypełniania infoboksów w różnych wersjach językowych Wikipedii. Nie ma ogólnie przyjętego standardu wypełniania atrybutów infoboksów. Dotyczy to m.in. niespójnych nazw atrybutów, zbioru używanych atrybutów czy też częstości wypełniania określonych atrybutów.

W przyszłości do analizy włączone zostaną dodatkowe miary jakości (związane z aktualnością, wiarygodnością), które mogą pomóc w weryfikacji danych w infoboksach przed przeniesieniem ich do innych wersji językowych. Użycie specjalnych metod może również zidentyfikować lepszą wersję językową, z której dane mogą wzbogacić mniej rozwinięte wersje (Lewoniewski, 2017). Na przykład jeden z projektów (Infoboxes.net, b.d.) przy użyciu niektórych ważnych miar może porównywać jakość infoboksów w 55 różnych wersjach językowych. Zwiększenie rozpatrywanych miar może zwiększyć wydajność podobnych narzędzi.

Interesujące są badania związane z analizą poszczególnych atrybutów infoboksu (a nie tylko infoboksu jako całość). To może pozwolić na do-

bór dwóch i więcej wersji językowych, w których poszczególne atrybuty o najlepszej jakości będą wykorzystane do wzbogacenia innych narodowych wersji Wikipedii. W konsekwencji może to podwyższyć jakość takich semantycznych baz wiedzy, jak DBpedia. Planowane są badania wpływu różnych miar dotyczących treści artykułu na jakość infoboksu. Wydzielić należy również kierunek badań związany z analizą popytu na informację, a jej wpływ na jakość danych w Wikipedii.

Proponowane metody mogą przyczynić się do poprawy jakości danych w korporacyjnych systemach informacyjnych – coraz więcej firm używa serwisów typu wiki do stworzenia swoich tematycznych baz wiedzy. Korporacyjne bazy wiedzy mogą być również wzbogacane za pośrednictwem Wikipedii przy wstępnej identyfikacji danych o najlepszej jakości za pomocą proponowanych metod.

Bibliografia

- Abramowicz, W. (2008). *Filtrowanie informacji*. Poznań: Wydawnictwo Akademii Ekonomicznej w Poznaniu.
- Alexa. (2018). *Wikipedia.org Traffic Statistics*. Pobrane 15 maja 2018 z <https://www.alexa.com/siteinfo/wikipedia.org>
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. i Hellmann, S. (2009). DBpedia–A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 154-165. <https://doi.org/10.1016/j.websem.2009.07.002>
- DBpedia. (b.d.). *Strona główna projektu DBpedia*. Pobrane 15 maja 2018 z <https://wiki.dbpedia.org/>
- DBpedia Mappings. (b.d.). Pobrane 15 maja 2018 z http://mappings.dbpedia.org/index.php/Main_Page
- Debattista, J., Auer, S. i Lange, C. (2016). Luzzu – A framework for linked data quality assessment. *Journal of Data and Information Quality (JDIQ)*, 8(1), 4. <https://doi.org/10.1109/ICSC.2016.48>
- GitHub. (b.d.). *Lewoniewski*. Pobrane 15 maja 2018 z <https://github.com/lewoniewski>
- Grudzień, Ł. (2012). Koncepcja oceny jakości informacji o procesach w systemach zarządzania. *Konferencja IZIP Zakopane*. Pobrane z http://www.ptzp.org.pl/files/konferencje/kzz/artyk_pdf_2012/p057.pdf
- Heinrich, B. i Klier, M. (2015). Metric-based data quality assessment–Developing and evaluating a probability-based currency metric. *Decision Support Systems*, 72, 82-96. <https://doi.org/10.1016/j.dss.2015.02.009>

- Infoboxes.net. (b.d.). Strona główna projektu *Infoboxes.net*. Pobrane 15 maja 2018 z <http://infoboxes.net>
- Ingawale, M., Dutta, A., Roy, R. i Seetharaman, P. (2013). Network analysis of user generated content quality in Wikipedia. *Online Information Review*, 37(4), 602-619. <https://doi.org/10.1108/OIR-03-2011-0182>
- Johnston, M. (2013). *Wikipedia revenue Analysis: How a Wiki could make \$2.3B a year*. Pobrane 15 maja 2018 z <https://monetizepros.com/features/analysis-how-wikipedia-could-make-2-8-billion-in-annual-revenue/>
- Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R. i Zaveri, A. (2014). Test-driven evaluation of linked data quality. *Proceedings of the 23rd international conference on World Wide Web*, 747-758. ACM. <https://doi.org/10.1145/2566486.2568002>
- Krytyka Wikipedii. (b.d.). W: *Wikipedia*. Pobrane 15 maja 2018 z https://pl.wikipedia.org/wiki/Krytyka_Wikipedii
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... Bizer, C. (2015). DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167-195. <https://doi.org/10.3233/SW-140134>
- Lewoniewski, W. (2017). Enrichment of information in multilingual Wikipedia based on quality analysis. *Lecture Notes in Business Information Processing*, 303, 216-227. https://doi.org/10.1007/978-3-319-69023-0_19
- Lewoniewski, W., Węcel, K. i Abramowicz, W. (2015). Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii. *Prace Naukowe: Systemy wspomagania organizacji SWO 2015*, Uniwersytet Ekonomiczny w Katowicach, 133-154.
- Lewoniewski, W., Węcel, K. i Abramowicz, W. (2016). Quality and importance of Wikipedia articles in different languages. *Communications in Computer and Information Science*, 639, 613-624. https://doi.org/10.1007/978-3-319-46254-7_50
- Lewoniewski.info. (b.d.). *Diagram Venna*. Pobrane 15 maja 2018 z <http://data.lewoniewski.info/soep2018/>
- Mendes, P.N., Mühleisen, H. i Bizer, C. (2012). Sieve: Linked data quality assessment and fusion. *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 116-123. ACM. <https://doi.org/10.1145/2320765.2320803>
- Meta-Wiki. (2018). *List of Wikipedias*. Pobrane 15 maja 2018 z https://meta.wikimedia.org/wiki/List_of_Wikipedias
- Paulheim, H. i Bizer, C. (2014). Improving the quality of linked data using statistical distributions. *International Journal on Semantic Web and Information Systems*, 10(2), 63-86. <https://doi.org/10.4018/ijswis.2014040104>
- Porównanie wyróżnień artykułów. (b.d.). W: *Wikipedia*. Pobrane 15 maja 2018 z https://pl.wikipedia.org/wiki/Pomoc:Porównanie_wyróżnień_artykułów
- Wakefield, J. (2017). Burger King advert sabotaged on Wikipedia. Pobrane 15 maja 2018 z <http://www.bbc.com/news/technology-39589013>

- Warncke-Wang, M., Cosley, D. i Riedl, J. (2013). Tell me more: An actionable quality model for Wikipedia. *Proceedings of the 9th International Symposium on Open Collaboration*, 8. ACM. <https://doi.org/10.1145/2491055.2491063>
- Węcel, K. i Lewoniewski, W. (2015). Modelling the quality of attributes in Wikipedia infoboxes. *Lecture Notes in Business Information Processing*, 228, 308-320. https://doi.org/10.1007/978-3-319-26762-3_27
- Wikimedia Downloads. (2018). *Data downloads*. Pobrane 15 maja 2018 z <https://dumps.wikimedia.org/>
- Wikimedia Foundation. (2016). *Financial Statements*. Pobrane 15 maja 2018 z https://upload.wikimedia.org/wikipedia/foundation/4/43/Wikimedia_Foundation_Audit_Report_-_FY15-16.pdf
- WikiRank. (b.d.). *Strona główna projektu WikiRank*. Pobrane 15 maja 2018 z <https://wikirank.net>
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. i Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1), 63-93. <https://doi.org/10.3233/SW-150175>