Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 705–714
DOI: 10.15439/2022F259
ISSN 2300-5963 ACSIS, Vol. 30

Identifying Reliable Sources of Information about Companies in Multilingual Wikipedia

Włodzimierz Lewoniewski, Krzysztof Węcel, Witold Abramowicz
Department of Information Systems, Poznan University of Economics and Business
Al. Niepodleglosci 10, Poznan 61-875, Poland
Email: {wlodzimierz.lewoniewski, krzysztof.wecel, witold.abramowicz}@ue.poznan.pl

***Abstract*—For over 21 years Wikipedia has been edited by volunteers from all over the world. Such editors have different education, cultural background and competences. One of the core rules of Wikipedia says, that information in its articles should be based on reliable sources and Wikipedia readers must be able to verify particular facts in text. However, reliability is a subjective concept and a reputation of the same source can be assessed differently depending on a person (or group of persons), language and topic. So each language version of Wikipedia may have own rules or criteria on how the website must be assessed before it can be used as a source in references. At the same time, nowadays there are over 1 billion websites on the Internet and only few developed Wikipedia language versions contain non-exhaustive lists of popular websites with reliability assessment. Additionally, since reputation of the source can be changed during the time, such lists must be updated regularly.**

This study presents the result of identification of reliable sources of information based on the analysis of over 200 million references that were extracted from over 40 million Wikipedia articles. Using DBpedia and Wikidata we identified articles related to various kinds of companies and found the most important sources of information in this area. This also allows to compare differences of the source reliability between Wikipedia languages.

I. Introduction

INFORMATION presented in Wikipedia articles should be based on reliable sources [1]. The source can be understood as the work (book, paper etc.), author, publisher. Such sources must have a proper reputation, should present all majority and significant minority views on some piece of information. Following this rule ensures that readers of the article can be assured that each provided specific fact (piece of information or statement) comes from a published and reliable source. Hence, before adding any information (even if it is a generally accepted truth) to this online encyclopedia, Wikipedia volunteer editors (authors or users) need to ascertain whether the facts put forward in the article can be verified by other people, who read Wikipedia [2].

Few developed language versions of Wikipedia contain non-exhaustive list of sources whose reliability and use on Wikipedia are frequently discussed. Even the English Wikipedia (the largest chapter of the encyclopedia) has such general list with information on reliability for less than 400 websites [3]. Sometimes we can find such lists for specific topics (e.q. video games, films, new Wikipedia articles in English Wikipedia).

It could take a significant human effort to produce a more complete list of assessed internet sources - there are over billion websites available in the Internet [4], [5] and a lot of them can be considered as a source of information. So, it can be very challenging and time consuming task for Wikipedia volunteers to assess reliability of each source. Moreover, reputation of each website can change with time - hence, such lists must be updated regularly. Additional challenge - each source may have a different reliability score depending on topic and language version of Wikipedia.

More complete and updated list of reliable sources can be useful not only for Wikipedia editors, but also for readers of this popular encyclopedia. The aim of this study is to show some possibilities of automating this process by analyzing existing and accepted content with sources in Wikipedia articles about companies in different languages. This paper uses existing and new models for reliability and popularity assessment of websites. The results show that depending on models it is possible to find such important sources in selected Wikipedia languages. Additionally, we show how the assessment of same sources can vary depending on language of this encyclopedia.

II. Related Works

Researching the quality of Wikipedia content is a fairly developed topic in scientific works. As one of the key factors influencing the quality of Wikipedia articles is the presence of references, some studies focused on researching information sources. Some of works use the number of references to automatically assess quality of the information in Wikipedia [6], [7], [8]. Such important measures are implemented in different approaches for automatic quality assessment of Wikipedia articles (for example WikiRank [9]). References often contain external links (URL addresses) where cited information is placed. Such links in references can be assessed by indicating the degree to which these conform to their intended purpose [10]. Moreover, those links can be employed separately to assess quality of Wikipedia articles [11], [12].

Some of the studies focused on metadata analysis of the sources in Wikipedia references. One of the previous works used ISBN and DOI identifiers to unify the references and find the similarity of sources between various Wikipedia language editions [13]. It is increasingly common practice to include scientific sources in references of Wikipedia articles. [13], [14], [15], [16]. At the same time, it is worth noting that such references often link to open-access works [17] and recently published journal articles [18]. One of the studies devoted to the COVID-19-related scientific works cited in Wikipedia articles and found that information comes from about 2% of the scientific works published at that time [19].

News websites are also one of the most popular sources of the information in Wikipedia and there is a method for automatic suggestion of the news references for the selected piece of information [20]. Particularly popular are references about recent content or life events [21]. For example in case of information related to COVID-19 pandemic Wikipedia editors inclined to cite the latest scientific works and insert more recent information on to Wikipedia shortly after the publication of these works [19].

Previous relevant publication [15] to this paper proposed and implemented 10 models for sources evaluation in Wikipedia articles. Results of assessment are also implemented in online tool "BestRef" [22]. Such approaches uses features (or measures) that can be extracted from publicly available data (Wikimedia Downloads [23]), so anybody can use those models for different purposes. One of the recent studies [24] in addition to the proposed models included also a time dimension to show how importance of the given web source of information on COVID-19 pandemic can be changed over different months.

III. References Extraction

To be able to extract information about references we prepared own parser in Python and applied it on Wikimedia dumps with articles in HTML format [23]. Table I presents te general statistics of the extraction.

External links (or URL addressees) in references were used to indicate main address of the website. However, each web source can use different structure of URL addresses. For example, some of the websites can use subdomains for separate topics of information or news. Another example - some organizational units (e.q. departments) of the same company may post its information on separate subdomains of main organization. To detect which level of domain indicates the source this work uses the Public Suffix List, which is a cross-vendor initiative to provide an accurate list of domain name suffixes [25]. Figure 1 presents example of URL address at fourth level domain with indication of main website.

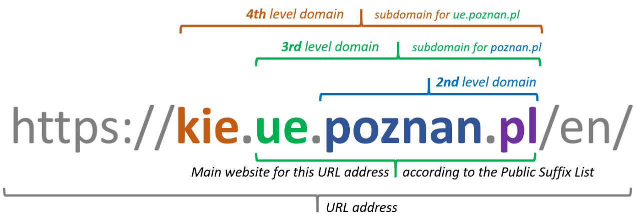

4th level domain
subdomain for ue.poznan.pl
3rd level domain
subdomain for poznan.pl
2nd level domain
https://kie.ue.poznan.pl/en/
Main website for this URL address
according to the Public Suffix List
URL address

Fig. 1. Example of URL address at fourth level domain with indication of main organizational website using the Public Suffix List

TABLE I
Statistics on references extraction from Wikipedia articles in different languages. Source: own calculations based on Wikimedia dumps in April 2022.

Abbr.	Language	Articles	References	Uniq. refs	RpA
ar	Arabic	1,162,992	6,689,241	5,208,058	5.75
be	Belarusian	216,747	589,402	453,54	2.72
bg	Bulgarian	280,546	935,65	727,127	3.34
ca	Catalan	698,608	3,350,195	2,637,219	4.80
cs	Czech	500,923	2,358,219	1,711,325	4.71
da	Danish	274,091	765,275	616,9	2.79
de	German	2,678,208	12,737,779	10,110,149	4.76
el	Greek	208,442	1,644,945	1,295,992	7.89
en	English	6,477,118	70,355,363	52,040,192	10.86
eo	Esperanto	315,637	302,146	257,393	0.96
es	Spanish	1,764,381	10,612,536	8,539,752	6.01
et	Estonian	226,552	548,589	419,373	2.42
eu	Basque	391,227	725,589	669,832	1.85
fa	Persian	892,984	2,012,489	1,748,880	2.25
fi	Finnish	528,323	2,938,331	1,916,372	5.56
fr	French	2,411,225	17,115,088	12,577,254	7.10
he	Hebrew	313,544	1,497,991	1,298,043	4.78
hr	Croatian	211,239	550,038	429,571	2.60
hu	Hungarian	501,758	2,241,596	1,646,175	4.47
hy	Armenian	291,266	1,853,522	1,294,452	6.36
id	Indonesian	618,676	2,170,068	1,700,961	3.51
it	Italian	1,748,062	7,769,065	5,780,364	4.44
ja	Japanese	1,319,693	12,153,736	8,237,546	9.21
kk	Kazakh	231,272	313,443	280,139	1.36
ko	Korean	584,594	1,599,714	1,327,504	2.74
lt	Lithuanian	202,444	486,654	447,025	2.40
ms	Malay	357,168	700,513	605,007	1.96
nl	Dutch	2,085,968	2,623,066	2,250,674	1.26
no	Norwegian (Bokmål)	582,399	1,874,697	1,490,498	3.22
pl	Polish	1,516,656	7,673,076	5,239,165	5.06
pt	Portuguese	1,088,286	6,636,422	5,116,972	6.10
ro	Romanian	428,682	2,021,351	1,327,598	4.72
ru	Russian	1,807,494	13,626,179	9,905,711	7.54
sh	Serbo-Croatian	456,444	1,368,842	909,406	3.00
simple	Simple English	207,354	630,729	515,962	3.04
sk	Slovak	240,027	562,559	456,986	2.34
sr	Serbian	657,077	3,234,971	1,760,098	4.92
sv	Swedish	2,580,001	11,695,159	7,875,678	4.53
tr	Turkish	477,885	2,216,325	1,567,293	4.64
uk	Ukrainian	1,146,175	4,291,799	3,457,589	3.74
vi	Vietnamese	1,271,057	3,392,140	2,846,216	2.67
zh	Chinese	1,264,023	6,730,567	5,182,993	5.32

Reference per Article (RpA) value shows average number of references in Wikipedia articles (in case of table I among separate language chapter). The highest value of this measure has English Wikipedia - almost 11 references per article. High values of RpA has also French (fr), Greek (el), Japanese (ja) and Russian (ru) Wikipedia.

IV. Models for Web Sources

Based on previous study [15], this work used following models for sources assessment with changes (described in this section):

1) **F**-model – how frequently (F) of considered source appears in references.
2) **PR**-model – how popular (P) are Wikipedia articles in which considered source appears divided by number of the references (R) in such articles.
3) **AR**-model – how much authors (A) edited the articles in which considered source appears divided by number of the references (R) in such articles.

One of the most basic and commonly used approaches to assess the importance of a web source is to count how frequently it was used in Wikipedia articles. This principle was used in relevant studies [26], [13], [27], [18]. So, **F**-model assesses how many times specific web domain occurs in external links of the references. For example, if the same source is cited 25 times in 13 Wikipedia articles (each contains at least one reference with such source), we count the (cumulative) frequency as 25. Equation 1 shows the calculation for F-model.

$$F(s) = \sum_{i=1}^{n} C_s(i), \tag{1}$$

where:

- s is the source, n is a number of the considered Wikipedia articles,
- $C_s(i)$ is a number of references using source s (e.q. domain in URL) in article i.

PR-model uses page views (or visits) of Wikipedia articles for certain period of time divided by the total number of all references in each considered Wikipedia article. Some studies showed correlation between information quality and page views in Wikipedia articles [28], [8], [29]. The more people read a specific Wikipedia article, the more likely its content was checked by part of them (including presence of reliable sources in references). So the more readers see the particular facts in the Wikipedia, the bigger probability that one of such reader will make appropriate edit if such facts are incorrect (or if source of information is inappropriate).

In other words, page views of the particular article usually shows the demand on information from Wikipedia readers. Therefore, visibility of the reference is also important. If more references are presented in the article, then the less visible is a specific source for the particular reader (visitor). At the same time, the more visitors has an Wikipedia article with references, the more visible is particular source in it. Equation 2 shows the calculation using PR-model.

$$PR(s) = \sum_{i=1}^{n} \frac{V(i)}{C(i)} \cdot C_s(i), \tag{2}$$

where:

- s is the source, n is a number of the considered Wikipedia articles,
- $C(i)$ is total number of the references in article i,
- $C_s(i)$ is a number of the references using source s (e.q. domain in URL) in article i,
- $V(i)$ is page views (visits) value of article for certain period of time i.

In comparison with previous research [15], for purposes of this study, apart from PR-model that uses cumulative page views V from humans (non-bots views) for a recent month (March 2022), additionally **PRy**-model will be used, which takes into account a wider date range - April 2021 - March 2022.

Quality of Wikipedia articles depends also on quantity and experience of authors who contributed to the content. Often articles in Wikipedia with the high quality are jointly created by a large number of different editors and this measure positively correlates with information quality [30], [31], [32], [33], [29]. To assess popularity of an article from editing users there is a possibility to analyze revision history of the article to find how many authors were involved in content creation/editing. So, AR-model shows how popular article is among Wikipedia volunteer editors. Equation 3 presents this model in mathematical form.

$$AR(s) = \sum_{i=1}^{n} \frac{E(i)}{C(i)} \cdot C_s(i), where: \tag{3}$$

- s is the source, n is a number of the considered Wikipedia articles,
- $C(i)$ is total number of the references in article i,
- $C_s(i)$ is a number of the references using source s (e.q. domain in URL) in article i,
- $E(i)$ is total number of authors of article i.

In contrast to previous work [15], AR-model in this study uses number of authors E that are registered on Wikipedia as users, without bot-users. Names of bots were selected based on the separate page (for example there is a special category in English Wikipedia [34]).

Additionally this study provides **ARe**-model, which is modification of AR-model:instead of counting the number of authors of a Wikipedia article, the number of editions of these authors (registered and non-bots) will be taken into account.

V. Using DBpedia and Wikidata to identify Wikipedia articles about companies

There are different possibilities to find topic of a particular Wikipedia article. For example, each article can be aligned to multiple categories, corresponding Wikidata item or DBpedia resource can highlight the topic based on properties in statements [29]. Additionally Wikipedia article can be included to different WikiProjects, that indicates interest to its information from groups of Wikipedia editors which focused on a specific topic (e.q. culture, history, military etc.).

This study used data from DBpedia and Wikidata to find Wikipedia articles related to companies. Each of those semantic databases have own advantages and disadvantages which are related to the operating principles and the technologies used.

A. DBpedia

DBpedia [35] is a semantic knowledge base that enriched automatically using structured information from Wikipedia articles in different languages [36], [37]. The resulting knowledge about some subject is available on the Web depending on title of Wikipedia article (as a source of that knowledge). For example, such semantic data about ”Meta Platforms” as the DBpedia resource we can find on the page https://dbpedia.org/resource/Meta_Platforms because such data were extracted from the relevant article in English Wikipedia - https://en.wikipedia.org/wiki/Meta_Platforms. At the same time DBpedia has separate knowledge extracted from other language versions and we can find also relevant information on other pages extracted from other Wikipedia chapters. On such DBpedia pages among different properties we can also find information about the type(s) of subject. In our example ”Meta Platforms” aligned to ”Company” and other classes of DBpedia ontology [38] and other structures. Such information is can generated automatically based on infoboxes (contained in Wikipedia articles) and their parameters. The figure 2 shows example of infoboxes about ”Meta Platforms” company in different Wikipedia languages. DBpedia extracts information about infoboxes based on the source code (wiki code or wiki markup) of the Wikipedia articles.

DBpedia ontology has a hierarchical structure, and if some resource is aligned to other company-related classes, we can use connections between those classes to detect Wikipedia articles related to companies. For example, some of the organizations can be aligned to ”Bank”, ”Publisher”, ”BusCompany” or other company-related class of DBpedia ontology, and after generalization we can find that all of them are belonging to ”Company” class. Based on DBpedia dumps related to instance types [35] (”specific” part of the dumps for each available language) we found that Wikipedia articles can be aligned directly to one of 634 classses from DBpedia ontology. Figure 3 shows those classes distinguishing with larger font size the most popular ones: *Person*, *Species*, *PopulatedPlace*, *Insect*, *Settlement*, *Place* and other. ”Company” class is the 20th most popular in such ranking.

It is worth mentioning that DBpedia provides two kinds of dumps that contain information on classification of resources (instances): instance-types (containing only direct types) and instance-types-transitive (containing the transitive types of a resource based on the DBpedia ontology). Such files contain triples of the form ’<*resource*> *rdf:type* <*class*>’ generated by the mappings extraction and other techniques for different language chapters of Wikipedia.

Figure 4 shows the structure of a part of DBpedia ontology with ”Organisation” class as a root node. It also presents information about directly alignments to separate classes of this ontology based on English Wikipedia. We can find there numbers based on of instances-types (direct alignment).

If we include also information on transitive types, we will have more resources aligned to same classes by taking into account connections between them in the DBpedia ontology. Figure 5 shows those classes distinguishing with larger font size the most popular ones: *Species*, *Eukaryote*, *Animal*, *Person*, *Location*, *Place* and other. ”Company” class is the 34th most popular in such ranking.

After considering transitive DBpedia dumps we have got additionally 20,736 resources (to directly aligned 64,372 resources) in ”Company” class - 85,108 in total in that class based on data from English Wikipedia. Next we took similar data extracted by DBpedia from other Wikipedia languages, and finally we got 173,418 unique companies [1]. Further we used ”DBO-companies” for the obtained list of Wikipedia articles about companies based on DBpedia extraction.

B. Wikidata

Wikidata [40] is a semantic knowledge base that works on a similar principles that Wikipedia with one important difference - here we can insert facts about the subjects using statements with properties and values rather then sentences in natural language. Wikidata is also considered as the central data management platform for Wikipedia and most of its sister projects [41].

Each Wikidata item has a collection of different statements structured in the form: ”Subject-Predicate-Object”. Figure 6 shows Wikidata item Q380 (”Meta Platforms”) with some statements.

Based on Wikidata statements we can find items on a specific topic. In our case, we will use the statement ”Property:P31 Q783794” (”instance of” - ”company”). Listing 1 presents SPARQL query to get such list from Wikidata using its query service [43]. Result of this query is available on the web page: https://w.wiki/5Bsc.

```
SELECT ?item WHERE {
        ?item wdt:P31 wd:Q783794. }
```

Listing 1. SPARQL query to get list of Wikidata items directly connected to ”company” item (Q783794) by ”instance of” property (P31)

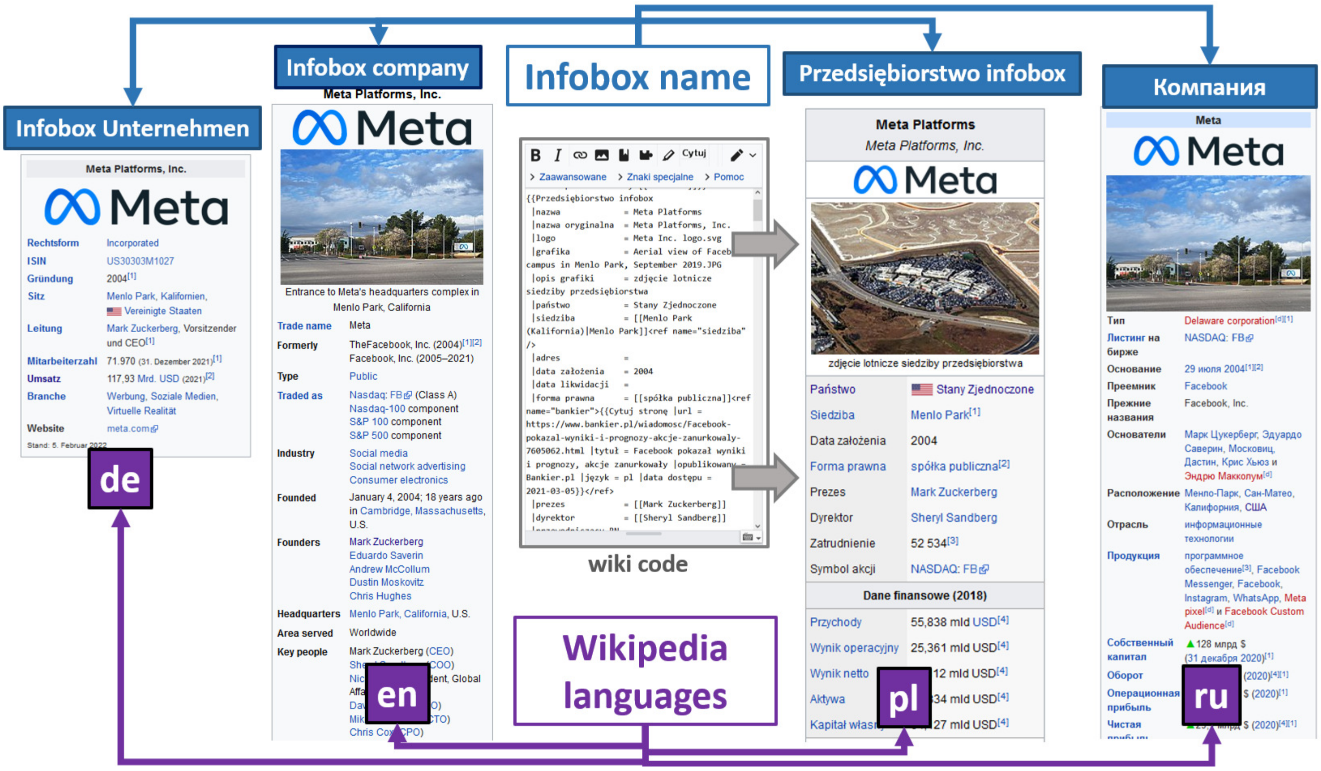
So, based on simple query we have got 12,635 Wikidata items related to companies. However, there are other connections in Wikidata that indicate items related to our topic. Similarly to DBpedia, here we can have also other ”sub-classes” or alternatives that can build more complete list of Wikidata items which can give list of appropriate Wikipedia articles. Let’s go back to our example on ”Meta Platforms” as an Wikidata item showed on the figure 6. We can see, that apart from ”company”, this item is also aligned to ”business” (Q4830453), ”enterprise” (Q6881511), ”public company” (Q891723) and ”technology company” (Q18388277) by ”instance of” parameter. Now we will use this information to enrich our query - listing 2 presents such SPARQL query: https://w.wiki/5Bsw. This query returned much more Wikidata items (comparing previous one) - 275,944 items. It is important to note, that this number doesn’t show directly number of Wikipedia articles related to companies, because not all Wikidata items contains links to at least one Wikipedia article.

```
SELECT ?item WHERE {
VALUES ?com {wd:Q783794 wd:Q4830453
 wd:Q6881511 wd:Q891723 wd:Q18388277}
?item wdt:P31 ?com.}
```

Listing 2. SPARQL query to get list of Wikidata items directly connected to ”company” (Q783794), business (Q4830453), enterprise (Q6881511), ”public company” (Q891723) and ”technology company” (Q18388277) by ”instance of” property (P31)

Despite significant increase of Wikidata items based on more complex query, there can be at least one important questions: is the proposed query complete enough to find all (or at least most of) Wikidata items related to companies?

[1] Unique company in this case means, that separate Wikipedia articles in various languages related to the same company counted as 1 company (instead of counting each Wikipedia article in each language version as a separate company).



Meta Platforms, Inc.

Rechtsform	Incorporated
ISIN	US30303M1027
Gründung	2004[1]
Sitz	Menlo Park, Kalifornien, Vereinigte Staaten
Leitung	Mark Zuckerberg, Vorsitzender und CEO[1]
Mitarbeiterzahl	71.970 (31. Dezember 2021)[1]
Umsatz	117,93 Mrd. USD (2021)[2]
Branche	Werbung, Soziale Medien, Virtuelle Realität
Website	meta.com

Stand: 5. Februar 2022

Meta Platforms, Inc.

Entrance to Meta's headquarters complex in Menlo Park, California

Trade name	Meta
Formerly	TheFacebook, Inc. (2004)[1][2] Facebook, Inc. (2005–2021)
Type	Public
Traded as	Nasdaq: FB (Class A) Nasdaq-100 component S&P 100 component S&P 500 component
Industry	Social media Social network advertising Consumer electronics
Founded	January 4, 2004; 18 years ago in Cambridge, Massachusetts, U.S.
Founders	Mark Zuckerberg Eduardo Saverin Andrew McCollum Dustin Moskovitz Chris Hughes
Headquarters	Menlo Park, California, U.S.
Area served	Worldwide
Key people	Mark Zuckerberg (CEO) She (COO) Nic dent, Global Affa Da O) Mik CTO) Chris Cox CPO)

B I Cytuj
Zaawansowane Znaki specjalne Pomoc

```
{{Przedsiębiorstwo infobox
 |nazwa              = Meta Platforms
 |nazwa oryginalna   = Meta Platforms, Inc.
 |logo               = Meta Inc. logo.svg
 |grafika            = Aerial view of Faceb
campus in Menlo Park, September 2019.JPG
 |opis grafiki       = zdjęcie lotnicze
siedziby przedsiębiorstwa
 |państwo            = Stany Zjednoczone
 |siedziba           = [[Menlo Park
(Kalifornia)|Menlo Park]]<ref name="siedziba"
/>
 |adres              =
 |data założenia     = 2004
 |data likwidacji    =
 |forma prawna       = [[spółka publiczna]]<ref
name="bankier">{{Cytuj stronę |url =
https://www.bankier.pl/wiadomosc/Facebook-
pokazal-wyniki-i-prognozy-akcje-zanurkowaly-
7605062.html |tytuł = Facebook pokazał wyniki
i prognozy, akcje zanurkowały |opublikowany =
Bankier.pl |język = pl |data dostępu =
2021-03-05}}</ref>
 |prezes             = [[Mark Zuckerberg]]
 |dyrektor           = [[Sheryl Sandberg]]
```

Meta Platforms
Meta Platforms, Inc.

zdjęcie lotnicze siedziby przedsiębiorstwa

Państwo	Stany Zjednoczone
Siedziba	Menlo Park[1]
Data założenia	2004
Forma prawna	spółka publiczna[2]
Prezes	Mark Zuckerberg
Dyrektor	Sheryl Sandberg
Zatrudnienie	52 534[3]
Symbol akcji	NASDAQ: FB
Dane finansowe (2018)	
Przychody	55,838 mld USD[4]
Wynik operacyjny	25,361 mld USD[4]
Wynik netto	12 mld USD[4]
Aktywa	34 mld USD[4]
Kapitał włas	27 mld USD[4]

Meta

Тип	Delaware corporation[d][1]
Листинг на бирже	NASDAQ: FB
Основание	29 июля 2004[1][2]
Преемник	Facebook
Прежние названия	Facebook, Inc.
Основатели	Марк Цукерберг, Эдуардо Саверин, Московиц, Дастин, Крис Хьюз и Эндрю Макколум[d]
Расположение	Менло-Парк, Сан-Матео, Калифорния, США
Отрасль	информационные технологии
Продукция	программное обеспечение[3], Facebook Messenger, Facebook, Instagram, WhatsApp, Meta pixel[d] и Facebook Custom Audience[d]
Собственный капитал	▲128 млрд $ (31 декабря 2020)[1]
Оборот	(2020)[4][1]
Операционная прибыль	$ (2020)[1]
Чистая	млрд $ (2020)[4][1]

Fig. 2. Infoboxes about "Meta Platforms" company in different Wikipedia chapters

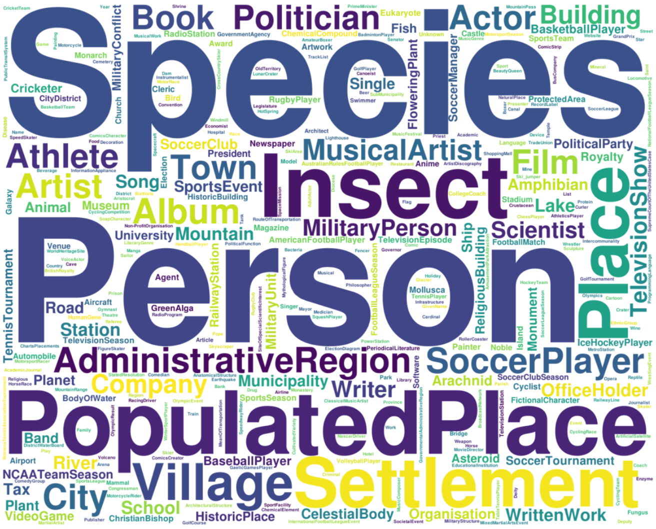


Fig. 3. Popular DBpedia ontology classes that are directly aligned to resources in various languages. Source: own calculations based on DBpedia ontology instance types specific dumps [35].

First, lets try to obtain general statistics on values that are inserted to "instance of" (P31) parameter among over 95 million Wikidata items. To do so, we prepared special algorithm in Python to extract such information from Wikidata dumps in JSON format [44]. It is worth noticing, that it is possible to construct SPARQL query to solve this task, however due to limitation of the Wikidata query service (such as limited time execution of the query) such statistics and other complex analysis can be done by extracting necessary data from the dump files. Figure 7 shows those items distinguishing with larger font size the most popular ones: *scholarly article* (Q13442814), *human* (Q5), *Wikimedia category* (Q4167836), *temporal range start* (Q523), *Taxon* (Q16521) , *infrared source* (Q67206691), *galaxy* (Q318) and other. Overall there are 87501 different alignments ("classes"). Items related to companies, such as "business" (Q4830453), "enterprise" (Q6881511) are on the 39th, 129th place respectively in such ranking.

Next we conduct such analysis only on Wikidata items, which has at least one link to Wikipedia article of one of the 42 considered languages in this study (see table I). Results are shown in figure 8. Now we have got 67,634 different alignments ("classes") and on the top we have: *Wikimedia category* (Q4167836), *human* (Q5), *taxon* (Q16521), *Wikimedia disambiguation page* (Q4167410), *Wikimedia template* (Q11266439), *human settlement* (Q486972), *Wikimedia list article* (Q13406463), *album* (Q482994), *film* (Q11424), *village* (Q532) and others. Early conidered items related to companies now are higher in the ranking: "business" (Q4830453) took 12th place, "enterprise" (Q6881511) took 66th place.

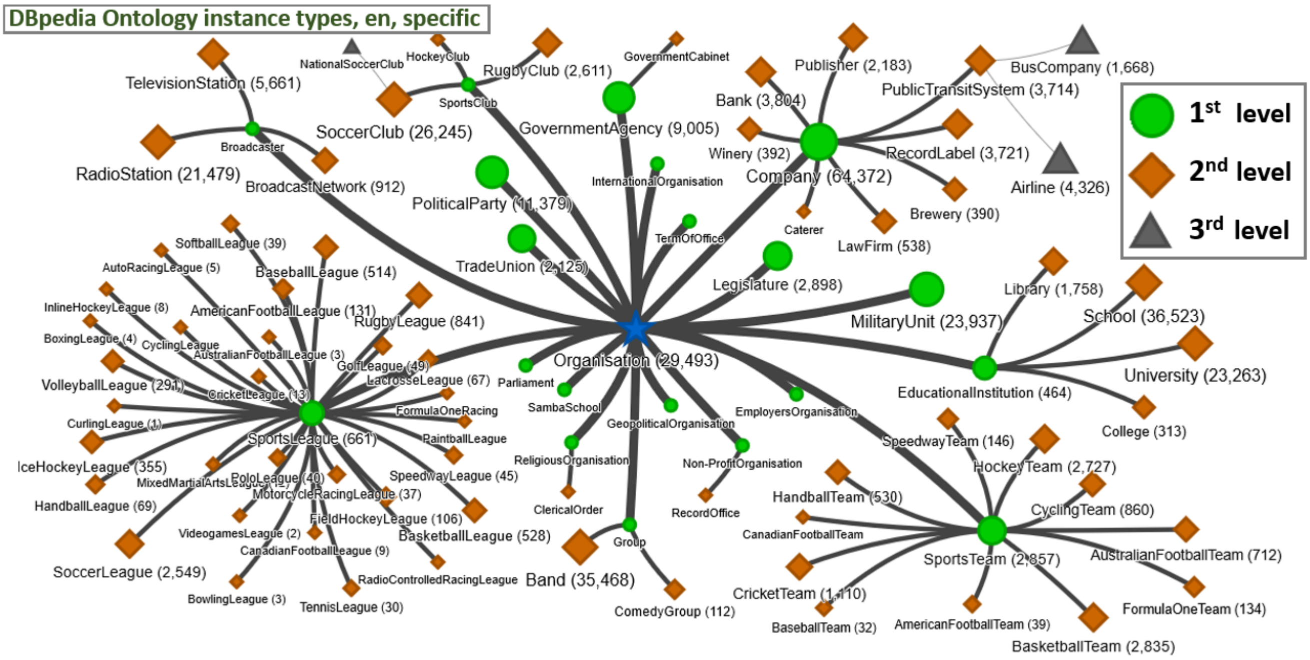


Fig. 4. Part of DBpedia ontology with "Organisation" as a root class. Number of articles from English Wikipedia aligned to a specific class of the ontology are given in brackets. Source: own calculations in April 2022 based on DBpedia dumps. Interactive version of the figure is available in [39]

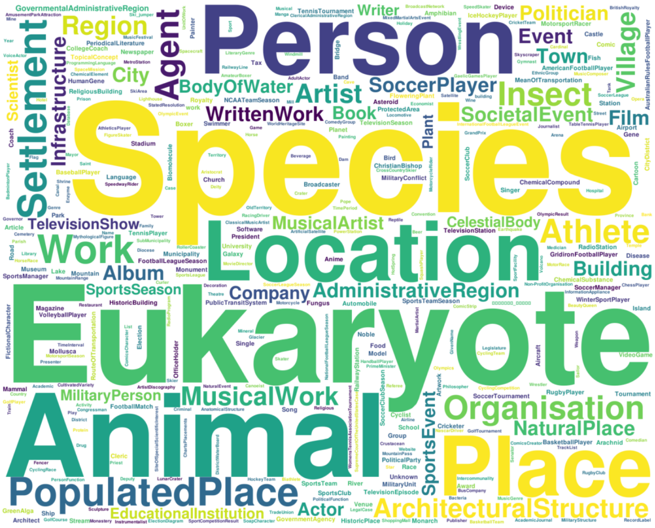
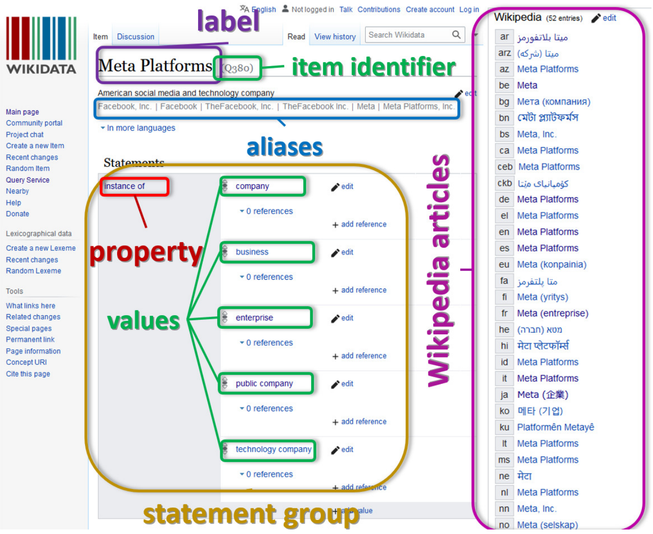


Fig. 5. Popular DBpedia ontology classes that are aligned to resources in various languages. Source: own calculations based on DBpedia ontology instance types (specific and transitive) dumps [35].



Meta Platforms (Q380)

American social media and technology company

Facebook, Inc. | Facebook | TheFacebook, Inc. | TheFacebook Inc. | Meta | Meta Platforms, Inc.

Statements

instance of

- company — 0 references
- business — 0 references
- enterprise — 0 references
- public company — 0 references
- technology company — 0 references

Wikipedia (52 entries)

ar	ميتا بلاتفورمز
arz	ميتا (شركه)
az	Meta Platforms
be	Meta
bg	Мета (компания)
bn	মেটা প্ল্যাটফর্মস
bs	Meta, Inc.
ca	Meta Platforms
ceb	Meta Platforms
ckb	کۆمپانیای مێتا
de	Meta Platforms
el	Meta Platforms
en	Meta Platforms
es	Meta Platforms
eu	Meta (konpainia)
fa	متا پلتفرمز
fi	Meta (yritys)
fr	Meta (entreprise)
he	מטא (חברה)
hi	मेटा प्लेटफॉर्म्स
id	Meta Platforms
it	Meta Platforms
ja	Meta (企業)
ko	메타 (기업)
ku	Platformên Metayê
lt	Meta Platforms
ms	Meta Platforms
ne	मेटा
nl	Meta Platforms
nn	Meta, Inc.
no	Meta (selskap)

Fig. 6. Scheme of the Wikidata item related to "Meta Platforms" company. Source: own work based on [42].

C. Combined approach

Comparing to DBpedia ontology classes (see V-A), Wikidata has much more possible aliments to different items - over 100 times more. To automatize process of identification company-related items in Wikidata there are various possibilities. One of them - to analyze Wikidata items related to "DBO-companies" selected using DBpedia extraction and find the most popular alignments in "instance of" statements. Figure 9 presents popular aliments for this case. Overall there are 3,453 various "classes" and the most popular are: business, enterprise, public company, company, automobile manufacturer, airline, record label, publisher, bus company, video game developer, organization, commercial organization, bank and others.

Finally let's take into the account alignments that appears

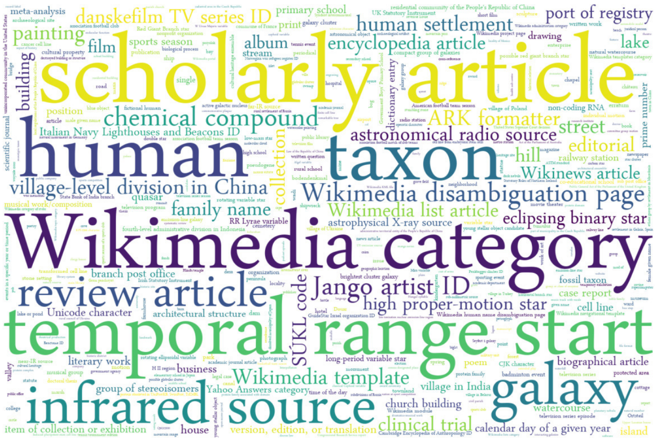


Fig. 7. Popular Wikidata items as a values in ”instance of” statements. Source: own calculations based on Wikidata dumps files [44].

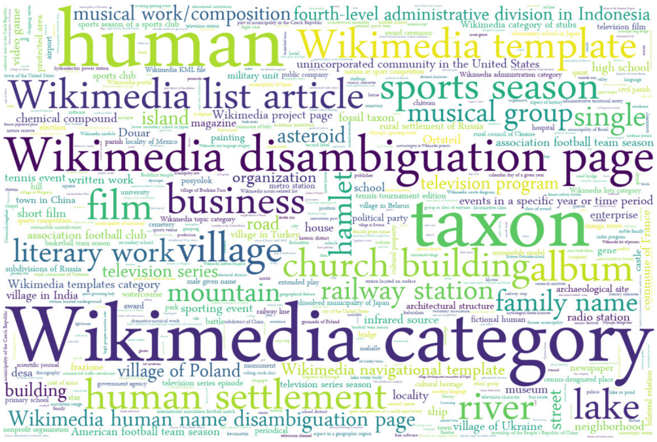


Fig. 8. Popular Wikidata items as a values in ”instance of” statements. Only Wikidata items with at least one link to Wikipedia article from one of 42 considered languages. Source: own calculations based on Wikidata dumps files [44].

at least 200 times to avoid insignificant mistakes that could be done by some users that edit Wikidata. In that case we will have 63 Wikidata items, that can appear in ”instance of” (P31) statements as a values. Additionally we removed alignment to ”organization” (Q43229) which is too general.

As a result, we have more Wikidata items with articles on the list of companies - overall 291,768 Wikidata items with at least one related Wikipedia article in considered language versions were identified. In futher analisys we will use ”WCA-companies” for this list.

VI. Estimating the information sources in Wikipedia about companies

This section presents results of assessment of the most important sources of information companies across Wikipedia languages using different models.

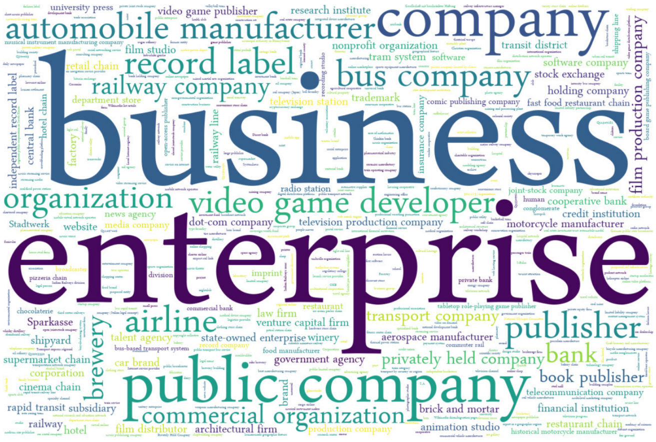


Fig. 9. Popular Wikidata items as a values in ”instance of” statements. Only Wikidata items with link to at least one Wikipedia article related to DBO-companies. Source: own calculations based on Wikidata dumps files [44].

Due to the limitation of space, following subsections presents results for the 15 most developed language versions of Wikipedia (with at least 1 million articles, see table I) Additionally, for the charts below, only the websites that appear at least 20 times in the top 100 at each language/model intersection[2] were selected. The more extended and interactive results can be found in supplementary materials [39].

It is important to note that archive services (such as archive.org) were excluded from analysis, due to the frequent occurrence of such links alongside the original sources in the same reference. If original source is no longer available, such archive services are very important, because Wikipedia readers can verify information, but unavailable original web sources are not a scope of this research. References to Wikipedia itself and Wikidata were also excluded. Links that are automatically inserted to references based on such identifiers as DOI (often links to doi.org) or ISBN (often links to books.google.com) cannot indicate directly the source of information. So such links were not considered in website analysis.

A. DBO-companies

First, we conducted a source analysis for the list of Wikipedia articles that have been generated based on data from DBpedia (see V-A) - ”DBO-companies”. Figure 10 shows the most important web sources of information on companies described in Wikipedia based with positions in rankings across 15 most developed language versions using five considered models.

Top 10 web sources in DBO-companies across 15 considered languages according to different models are as follows: nytimes.com, reuters.com, youtube.com, bloomberg.com, forbes.com, techcrunch.com, bbc.co.uk, cnn.com, wsj.com, theguardian.com.

[2] 15 languages and 5 models gives 75 such intersections

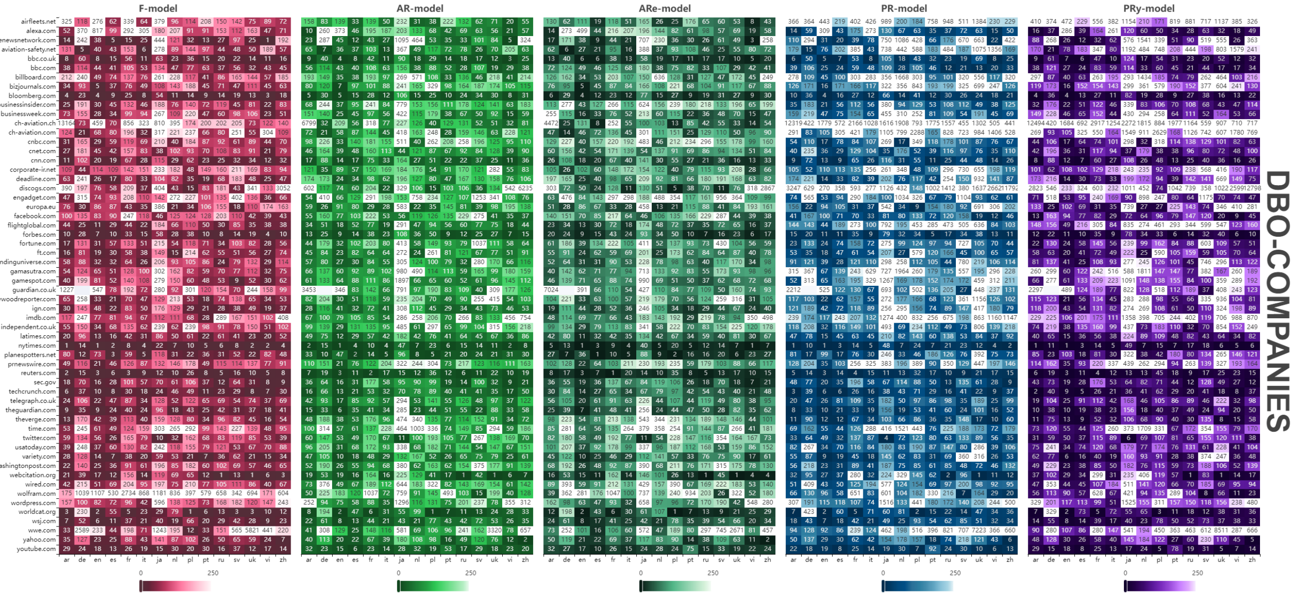


Fig. 10. The most important web sources of information on companies described in Wikipedia based on "DBO-companies" list with positions in rankings across 15 most developed language versions using various models. Source: own calculation based on Wikimedia dumps in April 2022. More extended and interactive version of the heat maps is available in [39]

B. WCA-companies

Figure 11 presents the most important web sources of information on companies described in Wikipedia based on WCA-companies list (described in V-B and V-C) with positions in rankings across 15 most developed language versions using five considered models.

Top 10 web sources in WCA-companies across 15 considered languages according to different models are as follow: nytimes.com, reuters.com, youtube.com, techcrunch.com, forbes.com, bloomberg.com, bbc.co.uk, theguardian.com, wsj.com, cnn.com.

C. Wikipedia languages

Based on average position in rankings calculated using different models we prepared the top 10 most important sources of information about companies in each Wikipedia languages.

Lists of such sources are presented below.

- **Arabic Wikipedia (ar)**: grid.ac, nytimes.com, worldcat.org, alexa.com, bbc.co.uk, bloomberg.com, reuters.com, techcrunch.com, theguardian.com, cnn.com
- **Belarusian Wikipedia (be)**: webcitation.org, tut.by, belta.by, zviazda.by, svaboda.org, nbrb.by, alexa.com, sec.gov, europa.eu, worldcat.org
- **Bulgarian Wikipedia (bg)**: capital.bg, brra.bg, dnevnik.bg, webcitation.org, alexa.com, bbc.co.uk, nytimes.com, forbes.com, vesti.bg, q4cdn.com
- **Catalan Wikipedia (ca)**: gencat.cat, elpais.com, worldcat.org, enciclopedia.cat, lavanguardia.com, ara.cat, vilaweb.cat, nytimes.com, elpuntavui.cat, elmundo.es
- **Czech Wikipedia (cs)**: idnes.cz, justice.cz, worldcat.org, ihned.cz, lupa.cz, novinky.cz, denik.cz, ceskatelevize.cz, e15.cz, zdopravy.cz
- **Danish Wikipedia (da)**: dr.dk, business.dk, politiken.dk, borsen.dk, finans.dk, computerworld.dk, berlingske.dk, tv2.dk, ing.dk, nytimes.com
- **German Wikipedia (de)**: spiegel.de, zdb-katalog.de, handelsblatt.com, mementoweb.org, heise.de, welt.de, faz.net, sueddeutsche.de, zeit.de, nytimes.com
- **Greek Wikipedia (el)**: et.gr, kathimerini.gr, tovima.gr, reuters.com, bbc.co.uk, capital.gr, nytimes.com, youtube.com, worldcat.org, typologies.gr
- **English Wikipedia (en)**: nytimes.com, worldcat.org, reuters.com, bbc.co.uk, bloomberg.com, theguardian.com, wsj.com, bizjournals.com, forbes.com, indiatimes.com
- **Esperanto Wikipedia (eo)**: staralliance.com, webcitation.org, liberafolio.org, wikimedia.org, wikiwix.com, nytimes.com, vortaro.net, debian.org, elpais.com, bloomberg.com
- **Spanish Wikipedia (es)**: elpais.com, issn.org, nytimes.com, elmundo.es, youtube.com, bbc.co.uk, lanacion.com.ar, planespotters.net, reuters.com, abc.es
- **Estonian Wikipedia (et)**: postimees.ee, err.ee, delfi.ee, riigiteataja.ee, aripaev.ee, muinas.ee, digar.ee, dv.ee, nasdaqbaltic.com, inforegister.ee
- **Basque Wikipedia (eu)**: berria.eus, worldcat.org, argia.eus, elpais.com, euskadi.net, euskadi.eus, eitb.eus, nih.gov, berria.info, diariovasco.com
- **Persian Wikipedia (fa)**: bbc.co.uk, bbc.com, webcitation.org, reuters.com, nytimes.com, sec.gov, forbes.com, alexa.com, isna.ir, radiofarda.com
- **Finnish Wikipedia (fi)**: yle.fi, hs.fi, kauppalehti.fi, is.fi, forbes.com, talouselama.fi, bloomberg.com, iltalehti.fi, taloussanomat.fi, nytimes.com
- **French Wikipedia (fr)**: lesechos.fr, lemonde.fr, reuters.com, lefigaro.fr, worldcat.org, societe.com, zonebourse.com, wikiwix.com, liberation.fr, lexpress.fr
- **Hebrew Wikipedia (he)**: globes.co.il, themarker.com, nli.org.il, ynet.co.il, calcalist.co.il, haaretz.co.il, walla.co.il, tase.co.il, mako.co.il, nytimes.com
- **Croatian Wikipedia (hr)**: bbc.co.uk, vecernji.hr, hrt.hr, zse.hr, tportal.hr, nytimes.com, enciklopedija.hr, jutarnji.hr, poslovni.hr, alexa.com
- **Hungarian Wikipedia (hu)**: index.hu, origo.hu, hvg.hu, youtube.com, nytimes.com, blog.hu, iho.hu, crt-tv.com, 24.hu, napi.hu
- **Armenian Wikipedia (hy)**: webcitation.org, nytimes.com, youtube.com, bbc.co.uk, sec.gov, purl.org, wsj.com, vedomosti.ru, kommersant.ru, forbes.com
- **Indonesian Wikipedia (id)**: detik.com, kompas.com, nytimes.com, forbes.com, worldcat.org, tempo.co, alexa.com, bbc.co.uk, reuters.com, kontan.co.id

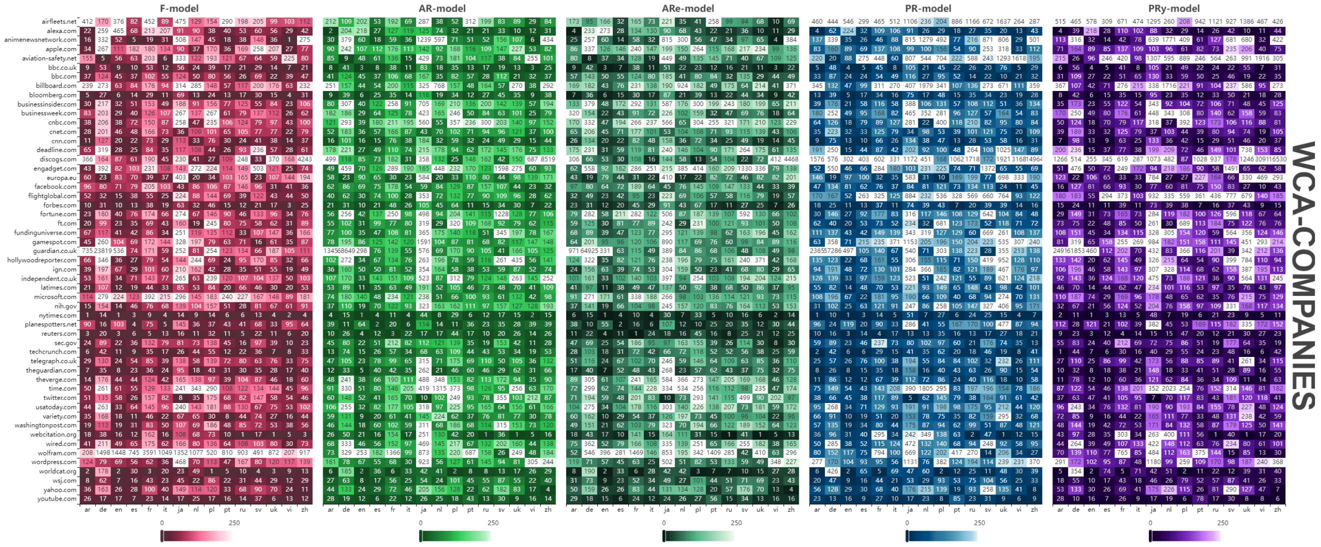


Fig. 11. The most important web sources of information on companies described in Wikipedia based on "WCA-companies" list with positions in rankings across 15 most developed language versions using various models. Source: own calculation based on Wikimedia dumps in April 2022. More extended and interactive version of the heat map is available in [39]

- **Italian Wikipedia (it)**: repubblica.it, corriere.it, ilsole24ore.com, nytimes.com, ansa.it, lastampa.it, bbc.co.uk, youtube.com, treccani.it, primaonline.it
- **Japanese Wikipedia (ja)**: catr.jp, nikkei.com, ndl.go.jp, impress.co.jp, asahi.com, itmedia.co.jp, twitter.com, eir-parts.net, edinet-fsa.go.jp, prtimes.jp
- **Kazakh Wikipedia (kk)**: webcitation.org, sec.gov, kase.kz, tengrinews.kz, bbc.co.uk, nytimes.com, lenta.ru, vedomosti.ru, shareholder.com, railways.kz
- **Korean Wikipedia (ko)**: naver.com, chosun.com, mt.co.kr, hankyung.com, mk.co.kr, donga.com, yonhapnews.co.kr, hani.co.kr, asiae.co.kr, khan.co.kr
- **Lithuanian Wikipedia (lt)**: vz.lt, delfi.lt, 15min.lt, vle.lt, bloomberg.com, lrytas.lt, ft.com, lrs.lt, lrt.lt, bbc.co.uk
- **Malay Wikipedia (ms)**: thestar.com.my, nytimes.com, bloomberg.com, sec.gov, utusan.com.my, forbes.com, reuters.com, worldcat.org, cnn.com, bbc.co.uk
- **Dutch Wikipedia (nl)**: nrc.nl, volkskrant.nl, nu.nl, nos.nl, fd.nl, standaard.be, telegraaf.nl, nytimes.com, ad.nl, kb.nl
- **Norwegian (Bokmål) Wikipedia (no)**: nb.no, nrk.no, brreg.no, e24.no, regjeringen.no, aftenposten.no, dn.no, snl.no, proff.no, vg.no
- **Polish Wikipedia (pl)**: virtualnemedia.pl, worldcat.org, wyborcza.pl, sejm.gov.pl, satkurier.pl, pwn.pl, rynek-kolejowy.pl, rp.pl, onet.pl, wp.pl
- **Portuguese Wikipedia (pt)**: uol.com.br, globo.com, abril.com.br, estadao.com.br, nytimes.com, worldcat.org, sapo.pt, forbes.com, terra.com.br, bloomberg.com
- **Romanian Wikipedia (ro)**: zf.ro, wall-street.ro, money.ro, adevarul.ro, capital.ro, mediafax.ro, evz.ro, hotnews.ro, nytimes.com, romanialibera.ro
- **Russian Wikipedia (ru)**: webcitation.org, kommersant.ru, vedomosti.ru, rbc.ru, lenta.ru, ria.ru, forbes.ru, tass.ru, reuters.com, cnews.ru
- **Serbo-Croatian Wikipedia (sh)**: nytimes.com, cnn.com, worldcat.org, bbc.co.uk, britannica.com, rts.rs, yahoo.com, washingtonpost.com, alexa.com, nih.gov
- **Simple English Wikipedia (simple)**: nytimes.com, wolfram.com, mathvault.ca, worldcat.org, bbc.co.uk, latimes.com, bloomberg.com, yahoo.com, reuters.com, sec.gov
- **Slovak Wikipedia (sk)**: worldcat.org, sme.sk, dennikn.sk, finstat.sk, etrend.sk, hnonline.sk, orsr.sk, aktuality.sk, pravda.sk, idnes.cz
- **Serbian Wikipedia (sr)**: b92.net, rts.rs, alexa.com, worldcat.org, nytimes.com, novosti.rs, politika.rs, apr.gov.rs, bbc.co.uk, blic.rs
- **Swedish Wikipedia (sv)**: allabolag.se, svd.se, dn.se, kb.se, svt.se, di.se, idg.se, mynewsdesk.com, worldcat.org, ne.se
- **Turkish Wikipedia (tr)**: hurriyet.com.tr, milliyet.com.tr, nytimes.com, haberturk.com, techcrunch.com, alexa.com, sec.gov, sabah.com.tr, youtube.com, bloomberg.com
- **Ukrainian Wikipedia (uk)**: webcitation.org, rada.gov.ua, rbc.ua, epravda.com.ua, pravda.com.ua, uprom.info, youtube.com, ukrinform.ua, nytimes.com, detector.media
- **Vietnamese Wikipedia (vi)**: nytimes.com, vnexpress.net, bbc.co.uk, tuoitre.vn, forbes.com, webcitation.org, bloomberg.com, youtube.com, techcrunch.com, animenewsnetwork.com
- **Chinese Wikipedia (zh)**: sina.com.cn, xinhuanet.com, qq.com, ltn.com.tw, yahoo.com, udn.com, sohu.com, chinatimes.com, nytimes.com, youtube.com

VII. Conclusion and Future Work

This study focused on information sources analysis of Wikipedia about companies in different languages. After extraction over 230 million references there were a process of indication of the main websites address for each URL address. As a result - over 2 million unique websites have been identified. To find important web sources across the languages, topics of the Wikipedia articles were analyzed. Using semantic representation of those information in DBpedia and user-generated knowledge in Wikidata this study shows how to find important web sources across languages based on existing and new models.

Models presented in this work can help not only Wikipedia volunteer editors to select web sites that can provide valuable information on companies, but also can help other Internet users to better understand how to find valuable sources of information a specific topic on the Web using open data from Wikipedia.

We plan to extend this research in future by providing additional features on identification of companies in Wikipedia. Additionally, we plan to divide different organizations to specific sectors (industries) to find the differences between reliability of information sources.

Future work will be focused also on extending reliability models and using different methods on topic classifications. One of the directions is to develop ways of weighting the importance of a reference based on its position within a Wikipedia article. There are also plans on including different measures related to the reputation of Wikipedia authors, protection of the articles, topic similarity and others.

Acknowledgements

The project financed within the Regional Initiative for Excellence programme of the Minister of Education and Science of Poland, years 2019-2023, grant no. 004/RID/2018/19, financing 3,000,000 PLN.

References

[1] English Wikipedia, "Wikipedia:Reliable sources," https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources, 2022.

[2] ——, "Wikipedia:Verifiability," https://en.wikipedia.org/wiki/Wikipedia:Verifiability, 2022.

[3] ——, "Wikipedia:Reliable sources/Perennial sources," https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources, 2022.

[4] Internet Live Stats, "Total number of Websites," https://www.internetlivestats.com/total-number-of-websites/, 2022.

[5] Netcraft, "August 2021 Web Server Survey," https://news.netcraft.com/archives/2021/08/25/august-2021-web-server-survey.html, 2021.

[6] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," *Proc. ICIQ*, pp. 442–454, 2005.

[7] J. E. Blumenstock, "Size matters: word count as a measure of quality on Wikipedia," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 1095–1096.

[8] W. Lewoniewski, "The method of comparing and enriching information in multlingual wikis based on the analysis of their quality," PhD, Poznań University of Economics and Business, 2018. [Online]. Available: http://www.wbc.poznan.pl/Content/461699/Lewoniewski_Wlodzimierz-rozprawa_doktorska.pdf

[9] WikiRank, "Quality and Popularity Assessment of Wikipedia Articles," https://wikirank.net/, 2022.

[10] P. Tzekou, S. Stamou, N. Kirtsis, and N. Zotos, "Quality Assessment of Wikipedia External Links," in *WEBIST*, 2011, pp. 248–254.

[11] E. Yaari, S. Baruchson-Arbib, and J. Bar-Ilan, "Information quality assessment of community generated content: A user study of Wikipedia," *Journal of Information Science*, vol. 37, no. 5, pp. 487–498, 2011.

[12] R. Conti, E. Marzini, A. Spognardi, I. Matteucci, P. Mori, and M. Petrocchi, "Maturity assessment of Wikipedia medical articles," in *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on*. IEEE, 2014, pp. 281–286.

[13] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Analysis of references across Wikipedia languages," in *International Conference on Information and Software Technologies*. Springer, 2017, pp. 561–573.

[14] F. Å. Nielsen, D. Mietchen, and E. Willighagen, "Scholia, scientometrics and Wikidata," in *European Semantic Web Conference*. Springer, 2017, pp. 237–259.

[15] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Modeling Popularity and Reliability of Sources in Multilingual Wikipedia," *Information*, vol. 11, no. 5, p. 263, 2020.

[16] H. Singh, R. West, and G. Colavizza, "Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia," *Quantitative Science Studies*, vol. 2, no. 1, pp. 1–19, 2021.

[17] M. Teplitskiy, G. Lu, and E. Duede, "Amplifying the impact of open access: Wikipedia and the diffusion of science," *Journal of the Association for Information Science and Technology*, vol. 68, no. 9, pp. 2116–2127, 2017.

[18] D. Jemielniak, G. Masukume, and M. Wilamowski, "The most influential medical journals according to Wikipedia: quantitative analysis," *Journal of medical Internet research*, vol. 21, no. 1, p. e11429, 2019.

[19] G. Colavizza, "COVID-19 research in Wikipedia," *Quantitative Science Studies*, vol. 1, no. 4, pp. 1349–1380, 12 2020. [Online]. Available: https://doi.org/10.1162/qss_a_00080

[20] B. Fetahu, K. Markert, W. Nejdl, and A. Anand, "Finding news citations for wikipedia," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 337–346.

[21] T. Piccardi, M. Redi, G. Colavizza, and R. West, "Quantifying engagement with citations on Wikipedia," in *Proceedings of The Web Conference 2020*, 2020, pp. 2365–2376.

[22] BestRef, "Popularity and Reliability Assessment of Wikipedia Sources," https://bestref.net, 2022.

[23] Wikimedia Downloads, "Main page," https://dumps.wikimedia.org, 2021.

[24] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Reliability in Time: Evaluating the Web Sources of Information on COVID-19 in Wikipedia across Various Language Editions from the Beginning of the Pandemic," 2022, presented at Wiki WorkShop 2022 (held virtually at The Web Conference 2022) on April 25, 2022.

[25] Public Suffix List, "List," https://publicsuffix.org/learn/, 2021.

[26] F. Å. Nielsen, "Scientific citations in Wikipedia," *arXiv preprint arXiv:0705.2106*, 2007.

[27] M. Redi, "Characterizing Wikipedia Citation Usage. Analyzing Reading Sessions," https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Citation_Usage/Analyzing_Reading_Sessions, 2019, [Online; accessed 01-Sep-2021].

[28] J. Lerner and A. Lomi, "Knowledge categorization affects popularity and quality of Wikipedia articles," *PloS one*, vol. 13, no. 1, p. e0190674, 2018.

[29] W. Lewoniewski, K. Węcel, and W. Abramowicz, "Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics," *Computers*, vol. 8, no. 3, 2019. [Online]. Available: https://www.mdpi.com/2073-431X/8/3/60

[30] A. Lih, "Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource," *5th International Symposium on Online Journalism*, p. 31, 2004.

[31] D. M. Wilkinson and B. a. Huberman, "Cooperation and quality in wikipedia," *Proceedings of the 2007 international symposium on Wikis WikiSym 07*, pp. 157–164, 2007.

[32] G. C. Kane, "A multimethod study of information quality in wiki collaboration," *ACM Transactions on Management Information Systems (TMIS)*, vol. 2, no. 1, p. 4, 2011.

[33] J. Liu and S. Ram, "Using big data and network analysis to understand Wikipedia article quality," *Data & Knowledge Engineering*, 2018.

[34] English Wikipedia, "Category:All Wikipedia bots," https://en.wikipedia.org/wiki/Category:All_Wikipedia_bots, 2022.

[35] Databus, "DBpedia Ontology instance types," https://databus.dbpedia.org/dbpedia/mappings/instance-types/, 2022.

[36] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[37] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, vol. 6, no. 2, pp. 167–195, 2015.

[38] DBpedia, "Ontology Classes," http://mappings.dbpedia.org/server/ontology/classes/, 2022.

[39] data.lewoniewski.info, "Supplementary materials for this research," https://data.lewoniewski.info/companies/, 2022.

[40] Wikidata, "Main Page," https://www.wikidata.org/wiki/Wikidata:Main_Page, 2022.

[41] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[42] Wikidata, "Q380," https://www.wikidata.org/wiki/Q380, 2022.

[43] Wikidata Query Sevice, "Main page," https://query.wikidata.org/, 2022.

[44] Wikimedia Downloads, "Wikidata Wiki Entities," https://dumps.wikimedia.org/wikidatawiki/entities/, 2022.