EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Qualitative assessment of Wikipedia as a source of big data on enterprises

Alexandros BITOULAS

WIH methodology team, Sogeti Luxembourg

Fernando Reis

Eurostat, WIH methodology team

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Web Intelligence Hub

Web Intelligence Hub (WIH)

- The **WIH** is the **pillar** of Trusted Smart Statistics that provides the fundamental building blocks for harvesting information from the web to produce statistics
- **Mission**: *"a high-quality source of data extracted from web content, methodologies and algorithms, ready to be used to produce European and national official statistics"*
- **Collaborative effort:** Eurostat, NSIs, statistical authorities and partners
- **Community of experts**: Web Intelligence Network, CEDEFOP
- **WIH Platform**: technical components and services
- **Current use cases**:
  - Online Job Advertisements
  - Online Based Enterprise Characteristics (OBEC)
  - Multinational Enterprises (MNE)

INSTITUTO NACIONAL DE ESTATÍSTICA | STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Outline

- Introduction
- Quality frameworks and Big Data
- UNECE Big Data Quality Framework (2014)
- Eurostat pipeline on enterprise data from Wikipedia
- Assessment of Wikipedia as a source of Big Data on Enterprises
- Conclusions, next steps

Wikipedia facts

Free online encyclopedia, hosted by Wikimedia Foundation

Created in 2001

World's largest reference website

Over 62 million articles in 300+ languages.

Open editing by volunteers

Potential source of enterprise data

Wikipedia

304 languages

Article Talk | Read View source View history Tools

From Wikipedia, the free encyclopedia

This article is about the online encyclopedia. For Wikipedia's home page, see Main Page. For the primary English-language Wikipedia, see English Wikipedia. For other uses, see Wikipedia (disambiguation).

Wikipedia[note 3] is a free content online encyclopedia written and maintained by a community of volunteers, known as Wikipedians, through open collaboration and the use of the wiki-based editing system MediaWiki. Wikipedia is the largest and most-read reference work in history.[3][4] It is consistently ranked as one of the ten most popular websites in the world, and as of 2024 is ranked the fifth most visited website on the Internet by Semrush,[5] and second by Ahrefs.[6] Founded by Jimmy Wales and Larry Sanger on January 15, 2001, Wikipedia is hosted by the Wikimedia Foundation, an American nonprofit organization that employs a staff of over 700 people.[7]

Initially only available in English, editions in other languages have been developed. Wikipedia's editions, when combined, comprise more than 63 million articles, attracting around 2 billion unique device visits per month and more than 14 million edits per month (about 5.2 edits per second on average) as of November 2023.[8][W 1] Roughly 26% of Wikipedia's traffic is from the United States, followed by Japan at 5.9%, the United Kingdom at 5.4%, Germany at 5%, Russia at 4.8%, and the remaining 54% split among other countries, according to data provided by Similarweb.[9]

Wikipedia has been praised for its enablement of the democratization of knowledge, extent of coverage, unique structure, and culture. It has been criticized for exhibiting systemic bias, particularly gender bias against women and geographical bias against the Global South (Eurocentrism).[10][11][*failed verification*] While the reliability of Wikipedia was frequently criticized in the 2000s, it has improved over time, receiving greater praise from the late 2010s onward,[3][10][12] while becoming an important fact-checking site.[13][14]

Wikipedia has been censored by some national governments, ranging from specific pages to the entire site.[15][16] Articles on breaking news are often accessed as sources for frequently updated information about those events.[17][18]

History

Main article: History of Wikipedia

Nupedia

Main article: Nupedia

Various collaborative online encyclopedias were attempted before the start of Wikipedia, but with limited success.[19] Wikipedia began as a complementary project for Nupedia, a free online English-language encyclopedia project whose articles were written by experts and reviewed under a formal process.[20] It was founded on March 9, 2000, under the ownership of Bomis, a web portal company. Its main figures were Bomis CEO Jimmy Wales and Larry Sanger, editor-in-chief for Nupedia and later Wikipedia.[1][21] Nupedia was initially licensed under its own Nupedia Open Content License, but before Wikipedia was founded, Nupedia switched to the GNU Free Documentation License at the urging of Richard Stallman.[W 2] Wales is credited with defining the goal of making a publicly editable encyclopedia,[22][W 3] while Sanger is credited with the strategy of using a wiki to reach that goal.[W 4] On January 10, 2001, Sanger proposed on the Nupedia mailing list to create a wiki as a "feeder" project for Nupedia.[W 5]

Launch and growth

The domains *wikipedia.org* and *wikipedia.com* (later redirecting to *wikipedia.org*) were registered on January 13, 2001,[W 6] and January 12, 2001,[W 7] respectively. Wikipedia was launched on January 15, 2001[20] as a single English-language edition at www.wikipedia.com[W 8] and was announced by Sanger on the Nupedia

Wikipedia

WIKIPEDIA
The Free Encyclopedia

The logo of Wikipedia, a globe featuring glyphs from various writing systems

Screenshot [show]

Type of site	Online encyclopedia
Available in	330 languages
Country of origin	United States
Owner	Wikimedia Foundation
Created by	Jimmy Wales Larry Sanger[1]
URL	wikipedia.org
Commercial	No
Registration	Optional[note 1]
Users	>287,987 active editors[note 2] >113,934,673 registered users
Launched	January 15, 2001 (23 years ago)
Current status	Active
Content license	CC Attribution / Share-Alike 4.0 Most text is also dual-licensed under GFDL; media licensing varies
Written in	LAMP platform[2]
OCLC number	52075003

Wikipedia founders Jimmy Wales (left) and Larry Sanger (right)

Wikipedia

304 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

This article is about the online encyclopedia. For Wikipedia's home page, see Main Page. For the primary English-language Wikipedia, see English Wikipedia. For other uses, see Wikipedia (disambiguation).

Wikipedia[note 3] is a free content online encyclopedia written and maintained by a community of volunteers, known as Wikipedians, through open collaboration and the use of the wiki-based editing system MediaWiki. Wikipedia is the largest and most-read reference work in history.[3][4] It is consistently ranked as one of the ten most popular websites in the world, and as of 2024 is ranked the fifth most visited website on the Internet by Semrush,[5] and second by Ahrefs.[6] Founded by Jimmy Wales and Larry Sanger on January 15, 2001, Wikipedia is hosted by the Wikimedia Foundation, an American nonprofit organization that employs a staff of over 700 people.[7]

Initially only available in English, editions in other languages have been developed. Wikipedia's editions, when combined, comprise more than 63 million articles, attracting around 2 billion unique device visits per month and more than 14 million edits per month (about 5.2 edits per second on average) as of November 2023.[8][W 1] Roughly 26% of Wikipedia's traffic is from the United States, followed by Japan at 5.9%, the United Kingdom at 5.4%, Germany at 5%, Russia at 4.8%, and the remaining 54% split among other countries, according to data provided by Similarweb.[9]

Wikipedia has been praised for its enablement of the democratization of knowledge, extent of coverage, unique structure, and culture. It has been criticized for exhibiting systemic bias, particularly gender bias against women and geographical bias against the Global South (Eurocentrism).[10][11][*failed verification*] While the reliability of Wikipedia was frequently criticized in the 2000s, it has improved over time, receiving greater praise from the late 2010s onward,[3][10][12] while becoming an important fact-checking site.[13][14]

Wikipedia has been censored by some national governments, ranging from specific pages to the entire site.[15][16] Articles on breaking news are often accessed as sources for frequently updated information about those events.[17][18]

Wikipedia

WIKIPEDIA
The Free Encyclopedia

The logo of Wikipedia, a globe featuring glyphs from various writing systems

Screenshot [show]

Type of site	Online encyclopedia
Available in	330 languages
Country of origin	United States
Owner	Wikimedia Foundation
Created by	Jimmy Wales Larry Sanger[1]
URL	wikipedia.org
Commercial	No
Registration	Optional[note 1]
Users	>287,987 active editors[note 2] >113,934,673 registered users
Launched	January 15, 2001 (23 years ago)
Current status	Active
Content license	CC Attribution / Share-Alike 4.0 Most text is also dual-licensed under GFDL; media licensing varies
Written in	LAMP platform[2]
OCLC number	52075003

History

Main article: History of Wikipedia

Nupedia

Main article: Nupedia

Various collaborative online encyclopedias were attempted before the start of Wikipedia, but with limited success.[19] Wikipedia began as a complementary project for Nupedia, a free online English-language encyclopedia project whose articles were written by experts and reviewed under a formal process.[20] It was founded on March 9, 2000, under the ownership of Bomis, a web portal company. Its main figures were Bomis CEO Jimmy Wales and Larry Sanger, editor-in-chief for Nupedia and later Wikipedia.[1][21] Nupedia was initially licensed under its own Nupedia Open Content License, but before Wikipedia was founded, Nupedia switched to the GNU Free Documentation License at the urging of Richard Stallman.[W 2] Wales is credited with defining the goal of making a publicly editable encyclopedia,[22][W 3] while Sanger is credited with the strategy of using a wiki to reach that goal.[W 4] On January 10, 2001, Sanger proposed on the Nupedia mailing list to create a wiki as a "feeder" project for Nupedia.[W 5]

Wikipedia founders Jimmy Wales (left) and Larry Sanger (right)

Launch and growth

The domains *wikipedia.org* and *wikipedia.com* (later redirecting to *wikipedia.org*) were registered on January 13, 2001,[W 6] and January 12, 2001,[W 7] respectively. Wikipedia was launched on January 15, 2001[20] as a single English-language edition at www.wikipedia.com [W 8] and was announced by Sanger on the Nupedia

Can be used as a source of data on Enterprises?

- Eurostat feasibility study of 2021
- NTTS 2023
- Wikipedia has a potential as a source of data on Enterprises
- Since 2022 developing a data collection on enterprise data from Wikipedia

The conference is partly financed by the European Union

Data Quality Frameworks

ESS Quality Framework:
Developed by Eurostat and NSIs

Big Data Quality Frameworks:

UNECE BDQF 2014: Technical approach, three phases: Input, Throughput, Output

The conference is partly financed by the European Union

ESS Quality Framework: principles and quality dimensions

Professional independence

Mandate for Data Collection and Access to Data

Adequacy of Resources

Commitment to Quality

Statistical Confidentiality and Data Protection

Impartiality and Objectivity

Sound Methodology

Appropriate Statistical Procedures

Non-excessive Burden on Respondents

Cost Effectiveness

Relevance

Accuracy and Reliability

Timeliness and Punctuality

Coherence and Comparability

Accessibility and Clarity

The conference is partly financed by the European Union

Big Data vs traditional data collections

Big data collected from non-traditional sources:

- from the web via scraping or via APIs
- generated by scanners (i.e. scanner or bar-code data)
- by mobile network operators (MNO data)
- from traffic cameras
- Etc.

The conference is partly financed by the European Union

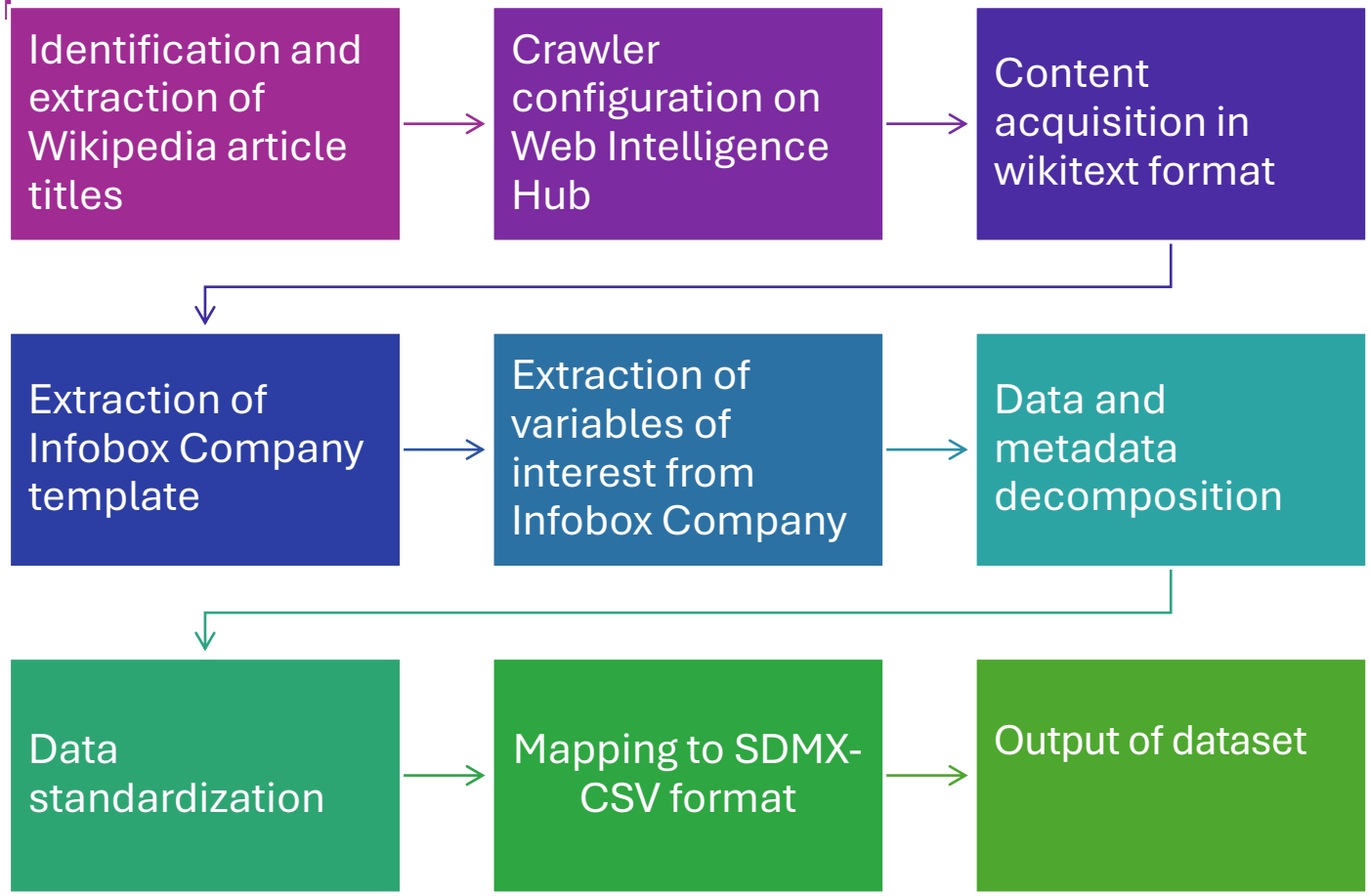
UNECE 2014 Big Data Quality Framework

- Developed by Statisticians for Statisticians, in 2014
- Approaches Big Data collections in a more appropriate, more “technical” way

The conference is partly financed by the European Union

UNECE 2014 Big Data Quality Framework

Hyperdimension / Phase	Input	Throughput	Output
Source	Institutional/Business Environment	System Independence	Institutional/Business Environment
	Privacy and Security	Steady States	Privacy and Security
Metadata	Complexity	Quality Gates	Complexity
	Completeness		Accessibility and Clarity
	Usability		Relevance
	Time-related factors		
	Linkability		
	Coherence-consistency		
	Validity		
Data	Accuracy and selectivity		Accuracy and selectivity
	Linkability		Linkability
	Coherence-consistency		Coherence-consistency
	Validity		Validity
	Usability		Time-related factors



Eurostat's Wikipedia Data Pipeline

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Input phase

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

The conference is partly financed by the European Union

Institutional Environment

This dimension refers to the institutional and organisational factors which may have a significant influence on the effectiveness and credibility of the source or of the agency producing the data.

•**Stability**:

- Part of the non-profit Wikimedia Foundation
- **More than sixty-two million articles** in **more than 300 languages**, including **6,809,459 articles in English**, with **123,762 active contributors in March 2024**
- Develops at a rate of over 2 edits every second and it averages 536 new articles per day
- Possiblity for a user to download the complete Wikipedia database in form of *dump files*. This allows for a user to keep archives or vintages of the database forever
- ***Can be considered a stable source***

The conference is partly financed by the European Union

Institutional Environment

- **Funding Model**:
  - Relies on public funding campaigns rather than advertisements
  - *Potentially ensuring a **certain degree of independence***
  - Potential sustainability risks if funding falls short.

- **Transparency**:
  - Any person can edit articles, with the history of edits (revisions) always available

The conference is partly financed by the European Union

Privacy and Security

This refers to the consent (active or passive) of the source to allow the scraping or downloading of its data, of whether physical

Passive Consent to download/extract data

Wikipedia allows data downloads via its API and dump files, subject to the conditions of the so called “robots.txt” policy

While formal consent was not successful to obtain, using the API was straightforward, without problems

The conference is partly financed by the European Union

Complexity

This dimension refers to the lack of simplicity and uniformity in the data structure including hierarchical complexity, the data format and the data source

- Data are generally unstructured
- With some exceptions: **Infobox templates**

Luckily for enterprise articles, Wikipedia has a specific template called ***'Infobox Company'***.

This template is a ***quasi-structured placeholder*** for demographic and economic variables for enterprises.

This provided a relatively structured, stable 'environment' of data for our data collection.

Complexity

Volkswagen AG (German: [ˈfɔlksˌvaːgn̩]), known internationally as the **Volkswagen Group**, is a German public multinational conglomerate manufacturer of passenger and commercial vehicles, motorcycles, engines and turbomachinery. Headquartered in Wolfsburg, Lower Saxony, Germany, and since the late 2000s is a publicly-traded family business owned by Porsche SE, which in turn is half-owned but fully controlled by the Austrian-German Porsche and Piëch family. The company also offers related services, including financing, leasing, and fleet management. In 2016, it was the world's largest automaker by sales, and keeping this title in 2017, 2018, and 2019, selling 10.9 million vehicles and was the largest automaker by revenue in 2022.[7] It has maintained the largest market share in Europe for over two decades.[8] It ranked seventh in the 2020 *Fortune* Global 500 list of the world's largest companies.[9] In Forbes Global 2000 2023 Volkswagen Group ranked 29th.[10]

The Volkswagen Group sells passenger cars under the Audi, Bentley, Cupra, Jetta, Lamborghini, Porsche, SEAT, Škoda and Volkswagen brands, motorcycles under the Ducati name, light commercial vehicles under the Volkswagen Commercial Vehicles brand, and heavy commercial vehicles via the marques of the listed subsidiary Traton (Navistar, MAN, Scania and Volkswagen Truck & Bus). It is divided into two primary divisions: the Automotive Division and the Financial Services Division. As of 2008, it had about 342 subsidiary companies.[11] Volkswagen also has three joint ventures in China, FAW-Volkswagen, SAIC Volkswagen and Volkswagen Anhui. The company has operations in roughly 150 countries, and it has 100 production facilities across 27 countries.

Volkswagen was founded in Berlin in 1937 and incorporated in Wolfsburg to manufacture the car that would become known as the Beetle. The company's production grew rapidly in the 1950s and 1960s. In 1965, it acquired Auto Union, which subsequently produced the first postwar Audi models. Volkswagen launched a new generation of front-wheel drive vehicles in the 1970s, including the Passat, Polo and Golf; the last became its bestseller. Volkswagen acquired a controlling stake in SEAT in 1986, making it the first non-German marque of the company, and acquired control of Škoda in 1994, of Bentley, Lamborghini, and Bugatti in 1998, Scania in 2008 and of Ducati, MAN, and Porsche in 2012. The company's operations in China have grown rapidly in the past decade, with the country becoming its largest market.

In 2015, Volkswagen was discovered to have used defeat devices to deceive environmental regulators about how much NO_x its cars were emitting. The company was fined billions of dollars.

Volkswagen Aktiengesellschaft is a public company and has a primary listing on the Frankfurt Stock Exchange, where it is a constituent of the Euro Stoxx 50 stock market index, and secondary listings on the Luxembourg Stock Exchange and SIX Swiss Exchange. It has been traded in the United States via American depositary receipts since 1988, currently on the OTC Marketplace. Volkswagen delisted from the London Stock Exchange in 2013.[12][13] The government of Lower Saxony holds 12.7% of the company's shares, granting it, by law, 20% of the voting rights.[14]

Volkswagen AG

VOLKSWAGEN GROUP

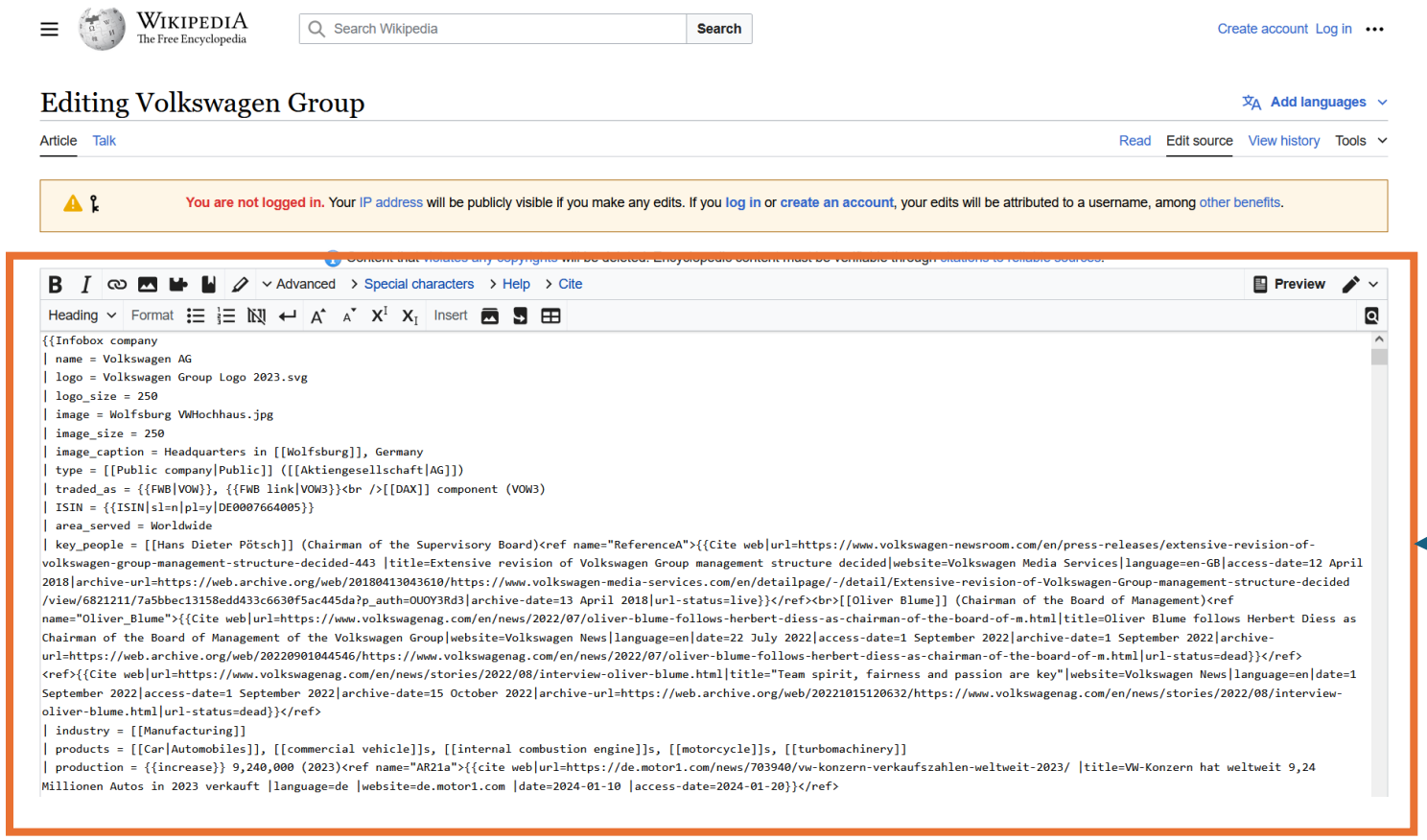
Headquarters in Wolfsburg, Germany

Company type	Public (AG)
Traded as	FWB: VOW, VOW3 DAX component (VOW3)
ISIN	DE0007664005
Industry	Manufacturing
Founded	28 May 1937; 87 years ago, in Berlin, Germany
Founder	German Labour Front
Headquarters	Wolfsburg, Lower Saxony, Germany
Number of locations	100 production facilities across 27 countries
Area served	Worldwide
Key people	Hans Dieter Pötsch (Chairman of the Supervisory Board)[1] Oliver Blume (Chairman of the Board of Management)[2][3]
Products	Automobiles, commercial vehicles, internal combustion engines, motorcycles, turbomachinery
Production output	▲ 9,240,000 (2023)[4]
Brands	**Automotive:**[5] [show] **Commercial:** [show] **Design:** [show]

Infobox company template, on a web browser

The conference is partly financed by the European Union

Complexity



Infobox company template, in wikitext format

Wikitext: Wikipedia's own markup language

The conference is partly financed by the European Union

Complexity

Overall ***medium*** complexity of the format of the data

For certain variables, (**ISIN**, **website**) parsing of data rather easy, due to relative 'standard' of formatting:

- domain.com for website variable
- 12 character alphanumeric code for ISIN

For economic variables, **net_income**, **assets**, web content not always followed expected format

In some other cases the parsing became rather more complex, as it required the development of ad-hoc regular expressions

The conference is partly financed by the European Union

Complexity

Wikipedia does not use standard codelists

- We developed dictionaries (ontologies of strings) to map all possible strings on Wikipedia to a specific Eurostat standard code
  - {Euro, Euros, Eur, euro, €} mapped to code 'EUR' of the Eurostat standard codelist 'CURRENCY'

- Few cases with **formatting and currency of numbers in non-western system**
  - Indian Rupees
  - values expressed in Canadian format

The conference is partly financed by the European Union

Completeness

This dimension refers to the extent to which metadata are available for a proper understanding and use of data

Wikipedia provides a basic definition of all variables of the Infobox Company template as well as a formatting template for users.

- Not of ESS standards
- *A form of structural metadata (i.e. metadata explaining data structure definition and record layout)?*

Editors of articles not always respect this recommended template - formatting

The conference is partly financed by the European Union

Completeness

For some variables, value is missing on Wikipedia (not edited by user), but fetched from Wikidata if exists there

- I.e. data available in html format but not in wikisource

The conference is partly financed by the European Union

Usability

This dimension refers to the extent to which we are able to work with and use the data without the employment of specialised resources or place significant burden on existing resources; and the ease with which it can be integrated with existing systems and standards.

Wikitext is Wikipedia's own markup language

Expert skills on **regular expression** to parse content from the infobox Company template

OpenSearch for data storage (NoSQL, json-like documents): potential training for a statistician to query and analyse this type of data

The conference is partly financed by the European Union

Timeliness and Periodicity

This refers to the added value of Big Data to be more timely and frequent than certain Official Statistics.

Timeliness of data from Wikipedia is ***mixed***

Early estimates for specific subpopulation of enterprises with available data on Wikipedia:

- For 40% of enterprises: data of reference year *T* updated with a delay of **T+3 months** after the reference period
- For 20% of enterprises: data refer to T-2 year reference period
- For a remainder 40% data were even older

More analysis required when more data are collected

A traditional data collection on enterprise timeliness would be at least T+12 months, potential for subgroups of population: e.g. large enterprises?

The conference is partly financed by the European Union

Accuracy (representativeness)

This refers to the degree to which the information correctly describes the phenomena it was designed to measure. Selectivity or representativeness refers to whether the information available on the Big Data Source differs from the information for the in-scope population

At this moment not possible to provide a quantitative assessment

Need to cross-check the data from Wikipedia against other reference datasets

The conference is partly financed by the European Union

Accuracy (representativeness)

However some qualitative aspects:

- Infobox company template used on approximately 85 000 articles
- 78 368 articles refer to 'company' (https://w.wiki/9jrW)
- 24 322 articles, articles referring to 'Enterprise' (https://w.wiki/9jrc) across all languages of Wikipedia

- I.e. a potential population of 85 000 -110 000 enterprise articles

The conference is partly financed by the European Union

Accuracy (representativeness)

- Selective in terms of coverage of the total population of enterprises globally
- May cover better specific sub-populations? Big enterprises, or multinational companies
    - Average number of monthly pageviews of the English Wikipedia article 'Volkswagen Group' = 120 000 during May 2023-April 2024
- **Financial reports are cited as reference for the economic variables:** a *proxy indicator of a high accuracy of information for these articles?*
- Further work needed in this direction

The conference is partly financed by the European Union

Coherence and linkability

Coherence: extent to which the dataset follows standard conventions, is internally consistent, is consistent over time and with other data sources.

Linkability, described as the ease with which the data can be linked or merged with other relevant datasets and consistency refers to the extent to which the dataset complies with standard definitions and is consistent over time.

Wikipedia definitions for economic and demographic variables seem to follow common principles and concepts.

- *Clear definitions: ISIN*, *headquarters location*, *website*
- *Potential evaluation by domain experts: revenue, net_income, assets, turnover* and *number_of_employees*

The conference is partly financed by the European Union

Coherence and linkability

- Presence of linking variables: *ISIN* or *website*
    - An assessment of the linkability of a specific sub-population of this dataset with EGR data was done by Eurostat in an earlier study (NTTS 2023)

The conference is partly financed by the European Union

Coherence and linkability

However:

- absence of ex-post validation of data at Wikipedia level
- no guarantee that values will be correct
- Processing errors in the form of data entry, typo or coding errors, or rounding issues may be found

Finally, yet no long time-series of this data is available, cannot evaluate *consistency over time,* in the future when more data will become available

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Throughput phase

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

The conference is partly financed by the European Union

Throughput phase

Refers to all intermediate stages between acquisition of the data and dissemination

3 important principles:

•**System independence**: the processing and transformation of the data should not be dependent on the system that is performing them

•**Steady States**: accessible intermediary versions of the dataset, which meet certain quality criteria

•**Quality Gates**: checkpoints in the statistical process at which the quality of the data is explicitly assessed

The conference is partly financed by the European Union

Throughput phase

Use of only open source tools: Apache Storm Crawler, OpenSearch database , R and Python

Early decision to use **steady states** to store the result of each step of the pipeline in the form of an easily accessible intermediate dataset

Evaluate any issues that may arise during the process and apply corrective actions if needed

The conference is partly financed by the European Union

Throughput phase

Area of improvement: absence of ***intermediary quality gates*** during the different steps of the process

Desing of a rather traditional (linear) approach of collecting, extracting and processing the data, leaving the validation of the data at the end of the process

Costs in resources between the different iterations and releases of the pipeline

- *To identify an error we had to run the complete pipeline and only evaluate the final output (dataset) for potential errors*
- *Lesson learnt and plans for refactoring the code base*

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Output phase

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

The conference is partly financed by the European Union

Output phase

*The quality of the output refers to the **reporting**, **dissemination** and **transparency** of the data and the process*

Privacy and confidentiality of the dataset: *Wikipedia's text are co-licenced under the Creative Commons Attribution-ShareAlike 4.0 International Licence (CC BY-SA) and the GNU Free Documentation Licence (GFDL).*

- Dataset produced by this pipeline can be accessible by any user

Complexity of the final dataset (output): no complexity in data structure/ format, as it follows and uses standard codelists and formats of SDMX

The conference is partly financed by the European Union

Output phase

At the moment no public formal documentation of the process is available.

Code base (Python) is available on Eurostat Gitlab

Process is being tested and refactored to meet higher standards of performance, efficiency and quality of the data before moved to PROD

Assessment of the accuracy and the selectivity of the data needs to be further carried out

Data can be collected whenever users need it

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Conclusions

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

The conference is partly financed by the European Union

Conclusions

Wikipedia, despite its open-editing nature, offers a rich and timely source of data that can enhance the statistical understanding of enterprises, particularly in terms of data timeliness

The **Input phase** highlighted Wikipedia's **stability** and **transparency** as a data source, albeit with challenges in **data complexity** and **standardization**

The **Throughput phase** emphasized the **need for quality gates in data processing** to maintain data integrity and reduce resource costs

Challenges: variability in data formats, need for advanced data processing techniques

Future work: evaluating accuracy, representativeness and consistency of Wikipedia-sourced data to enhance utility and reliability, move to prod for future users

The conference is partly financed by the European Union

Thank you

Stay connected

Alexandros.BITOULAS@ext.ec.europa.eu

Fernando.REIS@ec.Europa.eu

ESTAT-WIH@ec.europa.eu

https://ec.europa.eu/eurostat/web/main/home

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union