Qualitative Assessment of Wikipedia-Sourced Big Data On Enterprises

Alexandros BITOULAS[1], Fernando REIS[2]

[1]Sogeti, Luxembourg

[2]Eurostat, European Commission, Luxembourg

Abstract

In the evolving landscape of Big Data analytics, the integrity and quality of data are pivotal, especially in complex environments like enterprises. This study embarks on a qualitative assessment of Big Data on enterprises, with a unique focus on data sourced from Wikipedia, adhering to the suggested Big Data Quality Framework by UNECE (2014). Our analysis spans the initial stages of the data lifecycle - the Input and Throughput phases - and extends, to a lesser extent, to the Output phase. The Input phase examines the acquisition and pre-acquisition analysis of Wikipedia-sourced data, emphasizing aspects like the institutional and business environment, the complexity of the data, the completeness of metadata, the linkability and the selectivity of the data, all key components of the statistical quality of a data source. In the Throughput phase, we delve into the transformation, manipulation, and analysis of this data, underlining the principles of system independence, steady states, quality gates and discussing how the presence of unstructured information and noise can significantly influence the quality of the data. Additionally, an assessment of the Output phase is conducted, evaluating the reporting and dissemination qualities of the derived Big Data product, including its conformity to standards, coverage and overall relevance. By applying this comprehensive framework, we aim to provide an in-depth quality assessment that aligns with the intricate requirements of enterprises and underscores the value of Wikipedia as a source of Big Data for Official Statistics. Our findings are expected to contribute significantly to the discourse of the development of robust Big Data strategies in enterprises for Official Statistics, ensuring data quality and integrity throughout the statistical production process.

Keywords: Wikipedia, enterprises, Big Bata, quality

1 Introduction

Eurostat and its Web Intelligence Hub (WIH), is developing an innovative data collection on enterprise data, using data from publicly available web sources. One of the first sources of this data collection, is Wikipedia, which will be the focus of this article.

As per Wikipedia:

Wikipedia is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation. Written collaboratively by largely anonymous volunteers […], **Wikipedia articles can be edited by anyone** with Internet access (and who is not presently blocked), except in limited cases where editing is restricted to prevent

disruption or vandalism. Since its creation on January 15, 2001, it has grown into the world's largest reference website, attracting over a billion visitors monthly. Wikipedia currently has **more than sixty-two million articles in more than 300 languages**, including **6,809,459 articles in English**, with 123,762 active contributors in the past month (i.e. in March 2024). […]. Anyone can edit Wikipedia's text, references, and images. […] The content must conform with Wikipedia's policies, including being verifiable by published sources. Editors' opinions, beliefs, personal experiences, unreviewed research, libelous material, and copyright violations will not remain. Wikipedia's software allows easy reversal of errors, and experienced editors watch and patrol bad edits.

On the other hand:

Wikipedia makes **no guarantee of validity of the information** found here. The content of any given article may recently have been changed, vandalized, or altered by someone whose opinion does not correspond with the state of knowledge in the relevant fields [@wikipedi2024e].

Given the rich number of articles and information on companies or enterprises in general, Wikipedia may be used as a potential source of big data on enterprises. Following a recent feasibility study, Eurostat has decided to develop a pipeline to collect data on enterprises from Wikipedia. The lessons learnt and the experience gained so far are used to assess and evaluate the quality of Wikipedia as source of big data on enterprises.

The structure of this article is as follows:

- a short literature review on quality frameworks
- followed by the presentation of the Big Data Quality Framework of UNECE (2014), hereafter *UNECE_BDQF*, and its main dimensions.
- A presentation of the Eurostat data collection on enterprises from Wikipedia.
- This is then followed by the application of the quality assessment of Wikipedia as a source of Big Data, based on the specific dimensions of the *UNECE_BDQF*.
- Finally, a concluding section finishes this article.

2 Data Quality Frameworks

2.1 The European Statistical System (ESS) quality framework

Eurostat and the National Statistical Institutes (NSIs) of the EU Member States, altogether forming the European Statistical System (ESS), developed quite an extensive framework of

quality for Official Statistics [@eurostat_2021]. Together with it's **Code of Practice** [@european], Eurostat and the ESS define the following *16 principles of Official Statistics*:

1. **Professional independence** of statistical authorities from other policy, regulatory or administrative departments and bodies, as well as from private sector operators, ensures the credibility of European Statistics.
2. **Mandate for Data Collection and Access to Data:** Statistical authorities have a clear legal mandate to collect and access information from multiple data sources for European statistical purposes. Administrations, enterprises and households, and the public at large may be compelled by law to allow access to or deliver data for European statistical purposes at the request of statistical authorities.
3. **Adequacy of Resources**: The resources available to statistical authorities are sufficient to meet European Statistics requirements.
4. **Commitment to Quality:** Statistical authorities are committed to quality. They systematically and regularly identify strengths and weaknesses to continuously improve process and output quality.
5. **Statistical Confidentiality and Data Protection:** The privacy of data providers, the confidentiality of the information they provide, its use only for statistical purposes and the security of the data are absolutely guaranteed.
6. **Impartiality and Objectivity:** Statistical authorities develop, produce and disseminate European Statistics respecting scientific independence and in an objective, professional and transparent manner in which all users are treated equitably.
7. **Sound Methodology** underpins quality statistics. This requires adequate tools, procedures and expertise.
8. **Appropriate Statistical Procedures** implemented throughout the statistical processes, underpin quality statistics.
9. **Non-excessive Burden on Respondents**: The response burden is proportionate to the needs of the users and is not excessive for respondents. The statistical authorities monitor the response burden and set targets for its reduction over time.
10. **Cost Effectiveness:** Resources are used effectively.
11. **Relevance**: European Statistics meet the needs of users.
12. **Accuracy and Reliability:** European Statistics accurately and reliably portray reality.

13. **Timeliness and Punctuality**: European Statistics are released in a timely and punctual manner.
14. **Coherence and Comparability:** European Statistics are consistent internally, over time and comparable between regions and countries; it is possible to combine and make joint use of related data from different data sources.
15. **Accessibility and Clarity:** European Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.

The above framework of quality, which includes more technical indicators not presented here, is used to evaluate the content of most data collections carried out by the ESS. The main focus is on the two main types of *traditional* data collections, *surveys* and *administrative data*, plus the case of *Census*. While it's structure and concepts have been proven to fit the nature and properties of administrative data collections and surveys, they may not necessarily provide a sufficient fit for Big Data.

2.2 Big Data

Big Data are collected from non-traditional sources (i.e. surveys or administrative data). Instead they are collected from the web via scraping or via APIs, they can be generated by scanners (i.e. scanner or bar-code data), by mobile network operators (MNO data), from traffic cameras, among others.

They usually contain larger *variety* than traditional data collections and arrive in large *volumes* and with more *velocity* - some of the properties of Big Data, referred to also as the *three V's*.

When it comes to Big Data, although used extensively by their respective industries, e.g. MNO data used by telecommunication operators, scanner data generated by commercial or retail companies to monitor their sales, they only recently are being used in Official Statistics.

Some examples include *Statistics Netherlands,* with big data collections from traffic-loop data, social media messages and mobile phones [@netherlands_big_2020], *Statistics Denmark* which incorporated *bar code data* in the production of of the consumer price index from 1 January 2016 [@bigdata2018], *Eurostat* with data on skills and occupations from Online Job Advertisements [@labourm, @jobvaca].

2.3 Big Data Quality Frameworks

However, when it comes to quality frameworks for Big Data for Official Statistics, the available proposals are not many.

Eurostat and the ESS has not yet proposed a complete framework of qualitative assessment specific to Big Data. Therefore statisticians of Official Statistics of the ESS continue to either use the existing traditional framework of quality to assess the quality of the newest Big Data collections, or to use custom non-harmonised indicators of quality for Big Data.

As highlighted by [@reis2016], the *AAPOR Big Data Total Error framework* [@japec2015] extends the total survey error (TSE) to Big Data collections. This assumes the presence of *row errors*, *column errors* and *cell errors* and it illustrates three generic steps where errors may originate: generation, ETL (extraction, transform and load) and analysis.

2.4 UNECE Big Data Quality Framework (2014)

One of the deliverables of the UNECE Big Data Quality Task Team of 2014, was the *"Suggested Framework for the Quality of Big Data"* (*UNECE_BDQF*) [@unece_2014].

This framework approaches Big Data in a more "technical" way, compared to the Eurostat quality framework which was initially designed for surveys and administrative data.

Three distinct phases are proposed by the UNECE_BDQF, which split the overall business process of production of statistical outputs to:

- **Input**: when the data is acquired (collection stage);
- **Throughput**: any point in the process in which data is transformed or edited (processing phase, or ETL phase);
- **Output**: the assessment and reporting of quality of the statistical outputs derived from Big Data sources (evaluate and disseminate stage);

It uses a hierarchical structure composed of three **hyperdimensions,** with quality dimensions nested within each hyperdimension.

The three hyperdimensions are:

1. **Source**: which relates to factors such as the type of data, the source it was derived from as well as the governance under which the data is administered and regulated.
2. **Metadata**: which relates to the same factors as above.

3. **Data**: which relates to the quality of the data itself.

The complete matrix of the hyperdimensions and their respective dimensions are presented in table below, by phase of the statistical production process:

Table 1. Structure of dimensions of the UNECE_BDQF, by phase of the statistical process

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

One reason why we decided to focus on the UNECE_BDQF, is that it covers many of the dimensions of the Eurostat Quality framework and Code of Practice, while it also approaches Big Data collections in a way that it is more appropriate to the nature of Big Data, i.e. their more 'technical' nature, i.e. involving more intense IT routines and processes, compared to e.g. an administrative data collection or a survey.

3 Eurostat pipeline on enterprise data from Wikipedia

Our approach to extract content from Wikipedia involved a semi-automated algorithm. We developed scripts in R to identify relevant Wikipedia articles by exploring pertinent Wikipedia categories, such as "Companies by country", "Conglomerate companies of Germany" etc and we then acquired article titles within these categories using the Wikipedia API.

To extract (semi) structured data, we focused on Wikipedia's 'Infobox Company' template, utilized by approximately 85,000 articles, which contains demographic and economic information [@wikipedi2024f]. We accessed Wikipedia content through its API, in wikitext formats. This choice facilitated reliable and straightforward data extraction compared to

traditional web scraping and gave us access to the original source code (wikitext) Wikipedia articles are written to.

Our pipeline is designed in the following way including the following steps:

- Semi-automatic identification and extraction of titles of Wikipedia articles and their respective URLs
- creation of crawler on Eurostat's Web Intelligence Hub Platform [@webinte]. This step involved creating one crawler with all list of sources, i.e URLs of Wikipedia articles selected in previous step
- content acquisition in wikitext format
- extraction of the Infobox Company template
- Extraction of variables of interest
- Data and metadata decomposition
- Data standardization, i.e. tranformation of raw values to standard values using Eurostat standard codelists and formatting, e.g. for GEO, CURRENCY, numerical values etc.
- Mapping of the raw content (variables) to a final SDMX-CSV format and exporting of the SDMX-CSV file

The above pipeline was deployed in Python. The code is available for consultation on Eurostat's Gitlab [@webinte2024].

4 Application of the UNECE Big Data Quality Framework on Wikipedia

4.1 Input phase

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.

Wikipedia is part of the non-profit organisation Wikimedia Foundation. Since its creation on January 15, 2001, it has grown into the world's largest reference website, attracting over a billion visitors monthly. Wikipedia currently has 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 [@wikipedi2024d].

Wikipedia develops at a rate of over 2 edits every second and it averages 536 new articles per day [@wikipedi2024]. As can be seen in the table below, Wikipedia articles have been increasing across time [@wikipedi2024b]. Therefore the risk of Wikipedia being down or not existing in the future, can be assessed as low.

Table 2: Annual growth rate of articles for the English Wikipedia, source: Wikipedia

Date	Article count	Increase during preceding year	% Increase during preceding year
1/1/2002	19,700	19,700	—
1/1/2003	96,500	76,800	390%
1/1/2004	188,800	92,300	96%
1/1/2005	438,500	249,700	132%
1/1/2006	895,000	456,500	104%
1/1/2007	1,560,000	665,000	74%
1/1/2008	2,153,000	593,000	38%
1/1/2009	2,679,000	526,000	24%
1/1/2010	3,144,000	465,000	17%
1/1/2011	3,518,000	374,000	12%
1/1/2012	3,835,000	317,000	9%
1/1/2013	4,133,000	298,000	8%
1/1/2014	4,413,000	280,000	7%
1/1/2015	4,682,000	269,000	6%
1/1/2016	5,045,000	363,000	8%
1/1/2017	5,321,200	276,200	7%
1/1/2018	5,541,900	220,700	4.5%
1/1/2019	5,773,600	231,700	4.2%
1/1/2020	5,989,400	215,800	3.8%
1/1/2021	6,219,700	230,300	3.8%
1/1/2022	6,431,400	211,700	3.4%
1/1/2023	6,595,468	164,068	2.6%
1/1/2024	6,764,335	168,867	2.6%

Moreover, Wikipedia provides the possiblity for a user to download the complete Wikipedia database in form of *dump files* [@wikipedi2024a]. This allows for a user to keep archives or vintages of the database forever.

Another strong institutional parameter of Wikipedia is that it does not rely on advertisement for its income, but rather on public funcing campgaings [@wikipedi2024c]. This can assume a certain degree of independence. On the other hand this same issue could present a potential risk factor for the sustainability of the source for the future, i.e. in case Wikipedia cannot meet its needs of self-financing.

Finally, Wikipedia has very high standards on transparency. Any person may edit an article and the history of the editing process of an article is always registered so that any user can see the progress and history of a Wikipedia article. Articles can also go through a review process. The content must conform with Wikipedia's policies, including being verifiable by published sources. Editors' opinions, beliefs, personal experiences, unreviewed research, libelous material, and copyright violations will not remain. Wikipedia's software allows easy reversal of errors, and experienced editors watch and patrol bad edits [@wikipedi2024c].

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, technological and organisational provisions are in place to protect the security and integrity of statistical databases.

Wikipedia allows to download it's data in various formats, via it's API web service [@wikipedia_api]. Moreover users can download dump files of the complete database.

Before starting to download Wikipedia articles via the Wikipedia API, we tried to reach Wikipedia in advance for a formal consent to download its content. At first it was somewhat difficult to identify the appropriate person for such a correspondence e.g. the Data Owner or Data Controller but moreover we received no formal reply to our request. Due to the above and given the available information on Wikipedia, we assumed that we do not need any further active consent to use the API services, given we respect the API Etiquete conditions.

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.

When it comes to Wikipedia, data are generally unstructured. With some exceptions. Wikipedia uses specific *editing templates* for certain types of data, so that users can edit their content or values of certain variables based on a specific predefined template.

Luckily, for enterprise or company articles, Wikipedia has a specific template called *'Infobox Company'*. This template is a quasi-structured placeholder for demographic and economic variables for companies. This provided a relatively structured, stable 'environment' of data for our data collection.

An example of an *Infobox Company* template can be seen below:

```
{{Infobox company
| name = Volkswagen AG
```

```
| logo = Volkswagen Group Logo 2023.svg
| logo_size = 250px
| image = Wolfsburg VWHochhaus.jpg
| image_size = 250px
| image_caption = Headquarters in [[Wolfsburg]], Germany
| type = [[Public company|Public]] ([[Aktiengesellschaft|AG]])
| traded_as = {{FWB|VOW}}, {{FWB link|VOW3}}<br />[[DAX]] component (VOW3)
| ISIN = {{ISIN|sl=n|pl=y|DE0007664005}}
| area_served = Worldwide
| industry = [[Manufacturing]]
| revenue = {{increase}} [[Euro|€]]322.2840 billion (2023)
| operating_income = {{increase}} €21.586 billion (2023)
| net_income = {{increase}} €16.013 billion (2023)
| assets = {{increase}} €630.826 billion (2023)
| founder = [[German Labour Front]]
| location_city = [[Wolfsburg]], [[Lower Saxony]]
| location_country = Germany
| locations = [[list of Volkswagen Group factories|100 production facilities across 27 c
ountries]]
| homepage = {{URL|volkswagen-group.com}}
}}
```

Overall the complexity of the format of the data was medium. For certain variables, like ISIN and website (homepage in Wikipedia Infobox Company), parsing was relatively easy since the format of the data was following a relative 'standard' of formatting (e.g. domain.com for website variable and 12 character alphanumeric code for ISIN).

For economic variables, like net_income and assets, the web content did not always follow the expected format of e.g. '{{TREND}} CURENCYVALUE;UNIT (REF_YEAR)' as is seen in the above example (e.g. for assets = {{increase}} €630.826 billion (2023). In few cases of format discrepancy, the parsing became rather more complex, as it required the development of ad-hoc regular expressions.

Furthermore, Wikipedia does not require from editors to use or respect standard codelists when editing the values of certain variables. Therefore we had to develop certain dictionaries (or ontologies of strings) with all possible strings found on Wikipedia for a given code or class of data and to map them to a specific Eurostat standard code. For example strings {Euro, Euros, Eur, euro, €} were all mapped to code 'EUR' of the Eurostat standard codelist 'CURRENCY'.

In few other cases, similar complexities in parsing the values of economic variables came from the fact that for some articles the formatting and currency of numbers was in non-western system. This was the case for values in Indian Rupees or for values expressed in Canadian format.

Completeness

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

In terms of completeness of metadata, Wikipedia provides a guideline page of definitions of all variables of the Infobox Company template [@wikipedi2024g]. This is far from being a statistical definition of the standards of ESS, however it served as a main reference document explaining all the concepts for each variable of the Infobox Company template.

This minimum set of guidelines or recommendations on how editors must use the Infobox Company template and how they must format the values of each variable, could be parallerized with a form of *structural metadata* (i.e. metadata explaining data structure definition and record layout). Nevertherless, we observed that users (editors of Wikipedia articles) do not always respect this template and its recommended formatting.

Moreover, for some variables, if the value was missing, i.e. not edited by a user, but that same value existed in Wikidata, then this value was missing in the wikitext source code while it was available in the final html page of the Wikipedia article.

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. To do the first parsing of the infobox Company template, as well as to develop the regular expressions required for parsing most of the cases we identified, expert skills on Python and regular expression were necessary.

We used OpenSearch [@opensear] to store the original 'documents' (content from Wikipedia articles), as well as the intermediary data of the processing and extraction pipeline. This is a skillset for a very specific type of database, that normally not many statisticians possess and therefore certain training could be required for a statistician to be able to query and analyse this type of data.

Timeliness and Periodicity

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

We noticed that the timeliness of the data for the Wikipedia articles we examined was high. For many articles, the data for economic values, like net_income, revenue and assets for

reference year *T* was updated with a small delay of T+2 or T+3 months after the reference period, that is soon after when the financial results/reports were published by these enterrprises. This can be considered a strong quality aspect if we compare it with a traditional data collection on enterprise data, where normally data would require at least one year before they are collected by NSOs.

Accuracy (representativeness)

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

At this moment it is not possible to provide a quantitative assessment of the accuracy of the enterprise data collected from Wikipedia, neither of the representativeness or the selectivity of the population of enterprises on Wikipedia. For this type of assessment we would need to cross-check the data from the Wikipedia population against reference datasets.

However some qualitative aspects can already be highlighted. By querying the *Wikidata Query Service* about the number of Wikipedia articles referring to category 'company' (https://w.wiki/9jrW) we receive the result of 78 368 articles, while the same query for articles referring to 'Enterprise' (https://w.wiki/9jrc) returns a result of 24 322 articles, across all languages of Wikipedia. That means a potential population of more than 100 000 enterprises. While this dataset seems selective in terms of coverage of the total population of enterprises globally, it may be the case that it may cover better specific sub-populations. More specifically, following our experience so far, we were indeed able to identify quite a large number of big enterprises, i.e. enterprises with large presence in the global market, or multinational enterprises. This makes sense since Wikipedia is frequently used as means of communication of an enterprise to a specif target group of users, if we consider the number of daily visitors to Wikipedia articles. For illustrative purposes the average number of monthly pageviews of the English Wikipedia article 'Volkswagen Group' was more than 120 000 during the period May 2023-April 2024.

In many of the articles we analysed, financial reports of the enterprise were cited as reference for the economic variables. This can be perceived as a proxy indicator of a potentially high accuracy of information for these articles, as presumably the values for these variables were extracted from these official financial reports. However this assumption needs to be further tested and verified in the future.

Coherence and linkability

Coherence is the extent to which the dataset follows standard conventions, is internally consistent, is consistent over time and is consistent with other data sources. According to the UNECE_BDQF two subdimensions of coherence are important, the *linkability* of the data and the *consistency*.

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.

The definitions used for the economic and demographic variables of enterprises, as found in the documentation of the Infobox Company template, seem to follow common principles and concepts. For example variables like *ISIN*, *headquarters location*, *website*, have a straightforward definition. The same seems to apply for variables *revenue, net_income, assets* and *number_of_employees,* although for this latter group, expert's confirmation would be necessary as to whether definitions match fully with those used in international accounting standards.

Moreover, the presence of variables like *ISIN* or *website,* can be used as linking variables with other datasets of Official statistics.

Therefore, overall *linkability* of the Wikipedia dataset with other datasets (e.g. data from Business Registers) is potentially good. An assessment of the linkability of a specific sub-population of this dataset with EGR data was done by Eurostat in an earlier study [@conferen2021].

However the lack of very detailed *technical guidelines* of the variables avaialble on Wikipedia, as normally found in statistical data collections of Official Statistics, as well as the absence of ex-post validation of the values of enterprise data available on Wikipedia articles, provide no guarantee that the values will always be correct, that is they may suffer from processing errors in the form of data entry, typo or coding errors, or rounding issues.

Finally, at the time of writing, we are not able to have a long time-series of observations of this data, due to the recent nature of the project. Therefore we can not evaluate the consistency over time of this dataset at this stage. We plan to monitor and evaluate the consistency over time in the future when more data will become available.

4.2 Throughput phase

Throughput refers to all intermediate stages between acquisition of the data and dissemination. Three improtant principles of the UNECE_BDQF are described:

- **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.

Our pipeline is designed using only open source tools. We use Apache Storm Crawler for the crawler and OpenSearch database to store the web documents of the acquisition. We use R and Python to deploy the extraction and transformation process of the data. This makes our pipeline rather system-independent and could be easily adopted and reproduced by other users and still return the same results.

Moreover when designing the statistical pipeline, we took the 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. This resulted in 6 different and independent datasets during the process, starting from the first raw content of the Wikipedia article (in wikitext format), the extracted infobox, the filtered variables in-scope of the infobox, the transformed data and metadata, the standardised data and the eventual dissemination of the final SDMX-CSV dataset. This allows us to access each of this steady states datasets and evaluate any issues that may arise during the process and apply corrective actions if needed.

One area of improvement for our pipeline is the absence of intermediary quality gates during the different steps of the process. Our original idea was to follow a statistical approach of collecting, extracting and processing the data and leaving the validation of the data at the end of the process. That meant that we deployed a linear process of acquisition of the data (collection), extraction, transformation and loading (ETL) and finally dissemination of the dataset, without introducing any quality checks during the process. This resulted in 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. If an error would be discovered, the respective code would be amended and the pipeline had to be re-run again and the output re-evaluated for potential errors. We now learned from our

mistakes and we are now refactoring the code base to account for quality gates that will make the process easier to maintain as well as more efficient.

4.3 Output phase

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

In relation to the privacy and confidentiality of this dataset, *Wikipedia's text and many of its images are co-licenced under the Creative Commons Attribution-ShareAlike 4.0 International Licence (CC BY-SA) and the GNU Free Documentation Licence (GFDL). This means that Wikipedia content can be copied, modified and redistributed if and only if the copied version is made available on the same terms to others and acknowledgement of the authors of the Wikipedia article used is included* [@wikipedi2024h]. Therefore the eventual dataset produced by this pipeline can be accessible by any user, under the same terms.

The final dataset produced by our pipeline has no complexity with regard to data structure and format, as it follows and uses standard codelists and formats of SDMX.

At the moment there is no formal documentation and the process is being tested and refactored to meet higher standards of performance, efficiency and quality of the data, before it is moved to so called Production environment. Assessment of the accuracy and the selectivity of the data needs to be further carried out.

At the moment the data can be collected whenever needed, meaning enterprise data from Wikipedia may be available at the start of the year, where normally most enterprises release their financial results for reference year T-1. From a first analysis we observed that between January and March of year T most of the articles were updated with latest information on financial variables of reference year T-1. That means a timeliness of T+3 months after the end of the reference period. Of course a holistic assessment of the complete population needs to be undertaken on this aspect before generic conlusions can be drawn.

5 Conclusions

This study demonstrates the feasibility and value of using Wikipedia as a source of Big Data for enterprise statistics. By applying the UNECE Big Data Quality Framework, we evaluated the quality of Wikipedia-sourced data across the Input, Throughput, and Output phases. Our analysis revealed that 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 steady states and quality gates in data processing pipelines to maintain data integrity and reduce resource costs. Finally, the Output phase demonstrated the importance of adhering to privacy and licensing regulations, ensuring that the final dataset is both accessible and reliable.

In conclusion, while Wikipedia presents certain challenges, such as variability in data formats and the need for advanced data processing techniques, it also offers significant opportunities for enriching official statistics with up-to-date and diverse data. Our findings support the development of robust Big Data strategies that leverage unconventional sources like Wikipedia, ensuring high data quality and integrity throughout the statistical production process. Future work should focus on evaluating the accuracy and representativeness, as well as the consistency of Wikipedia-sourced data and to further integrate it with other official data sources to enhance its utility and reliability.

6 References

2014 Project - Big Data in Official Statistics - UNECE Statswiki. (n.d.). Retrieved April 18, 2024, from https://statswiki.unece.org/display/bigdata2/2014+Project?preview=/108102944/108298642/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf

API:Main page. (n.d.). https://www.mediawiki.org/wiki/API:Main_page

Big Data Strategy 2018-2020. (2018). Statistics Denmark. https://www.dst.dk/publ/BigDataStratUK

Conference on New Techniques and Technologies for Statistics. (2021). https://wayback.archive-it.org/12090/20231227180744/https://cros-legacy.ec.europa.eu/content/NTTS2023_en

European Commission. Statistical Office of the European Union. (2021). *European Statistical System handbook for quality and metadata reporting: 2021 re edition.* Publications Office. https://data.europa.eu/doi/10.2785/616374

European statistics code of practice - eurostat. (n.d.). https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880. https://doi.org/10.1093/poq/nfv039

Job vacancy statistics. (n.d.). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Job_vacancy_statistics

Labour market demand for ICT specialists in online job advertisements - eurostat. (n.d.). https://ec.europa.eu/eurostat/web/experimental-statistics/labour-market-demand-ict-specialists

Netherlands, S., Braaksma, B., & Zeelenberg, K. (2020). *Big data in official statistics* [Webpagina]. https://www.cbs.nl/en-gb/background/2020/04/big-data-in-official-statistics

OpenSearch. (n.d.). https://opensearch.org/

Reis, F., Di Consiglio, L., Kovachev, B., Wirthmann, A., & Skaliotis, M. (2016). *Comparative assessment of three quality frameworks for statistics derived from big data: The cases of wikipedia page views*

and automatic identification systems. *31*. https://scholar.google.com/scholar?cluster=8999163208076034068&hl=en&oi=scholarr

Web intelligence hub, Eurostat. . https://cros.ec.europa.eu/landing-page/web-intelligence-hub

Web Intelligence Hub Platform · GitLab. (2024). https://git.fpfis.tech.ec.europa.eu/estat/wihp

Wikipedia Template: Infobox company. (2024a). https://en.wikipedia.org/w/index.php?title=Template:Infobox_company/doc&oldid=1220303290

Wikipedia:About. (2024a). https://en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=1221438217

Wikipedia:Copyrights. (2024c). https://en.wikipedia.org/w/index.php?title=Wikipedia:Copyrights&oldid=1216438911

Wikipedia:Funding Wikipedia through advertisements. (2024d). https://en.wikipedia.org/w/index.php?title=Wikipedia:Funding_Wikipedia_through_advertisements&oldid=1222219149

Wikipedia:Size of Wikipedia. (2024e). https://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=1223114961

Wikipedia:Statistics. (2024g). https://en.wikipedia.org/w/index.php?title=Wikipedia:Statistics&oldid=1222828616