Open data sources for retrieving information on multinational enterprise groups

**Keywords:** Statistical Business Register, EuroGroups Register (EGR), Enterprise, aggregated group figures, crowdsourcing, open data projects.

# Introduction

Eurostat governs the EuroGroups Register (EGR), the statistical business register of multinational enterprise groups (MNEs) in the European Union and EFTA countries. Sources of information such as crowdsourcing platforms, web crawling and different open data projects are seen as further opportunities to increase the quality of the EGR, its completeness and accuracy namely with the units outside of the EU and EFTA as well as on the whole group level.

Under the umbrella of Eurostat BIG DATA project, the EGR Team is investigating these additional data sources. Eurostat is collaborating with Leipzig University to explore the possibility of using DBpedia as new additional source of data of multinational enterprise groups. EGR Team and Leipzig University carried out a Prove of concept (PoC) to automatize at a large extend the collection of aggregated group figures using as input the names of the enterprise groups.

# Methods

Currently EGR Team performs final checks and updates of enterprise group figures manually by looking into the data published in the annual accounts or websites. For the feasibility study EGR team selected 73 enterprise groups (EGs) of reference year 2016 based on size and geographical diversity and provided them to DBpedia for matching. The main attributes, which were targeted to be collected, were persons employed, turnover and assets. The following indicators were analysed:

* Coverage - number of successfully matched enterprise groups’ names
* Completeness - number of received values for the different attributes
* Accuracy - quality of the returned values when compared to the figures published by the enterprise itself
* Timeliness - availability of data for certain reference period based on EGR cycle

DBpedia is a project that extracts data from Wikipedia (info boxes) in order to make it publicly available in the machine readable and query-friendly Resource Description Framework (RDF) format that overcomes limitations of the latter. Wikipedia is an online encyclopaedia that is collaboratively written by volunteers. Currently about 71000 active contributors are working on more than 47 Million articles in 299 languages. Its specific way of content creation brings advantages and disadvantages. It is a very rich data source but does not provide any guarantee for the quality of the data. In addition, the structure and representation of the information is quite diverse, especially across different language versions.

However, in which way and to what extend DBpedia can be considered a potential resource to enrich the EGR Database on enterprise groups does not only depend on the data source itself, but also on the technical possibilities to create a practical tool for parsing data. The project goal was to create a Prove of Concept as an interface that handles a list of enterprise names and returns a list of results with detailed information on those enterprises. The workflow of this tool is split into 3 phases depending on each other: Linking (company Identification), Data Download & Extraction as well as a Data Conversion, Consolidation & Export phase.

## Phase 1: Entity Linking (MNE Identification)

Every MNE name is sent to the DBpedia Spotlight Application program interface (API) in order to identify and annotate it with a DBpedia ID (which corresponds to a Wikipedia Page). Multiple language models are used which improves the recognition of language dependent aliases or name variations and also improves coverage for entities which have only a restricted regional importance. Spotlight tries to disambiguate names but it can deliver multiple candidates with individual confidence scores.

## Phase 2: Data Download and Extraction

Using the Wikimedia API plus the DBpedia ID resolved in phase 1, all available Wikipedia articles for set of (currently 17 European) languages are retrieved for an MNE. The DBpedia extraction framework is used to extract and transform the information primarily contained in the info boxes of the Wikipedia articles into RDF. Reference year information from e.g. employment figures and revenues very often are not provided for the same reference year. Furthermore, data for the year 2016 tends to be not available for bigger companies in Q2/2018 anymore, since Wikipedia users also updated the article for 2017. In order to achieve better coverage for a given reference year, a historic mode is integrated in the prototype. To put it briefly, this mode tracks all changes in the info box and merges this information into a virtual file per reference year. Such a file represents all information for one entity from a single Wikipedia language version, which could be retrieved for the target reference year.

## Phase 3: Data Conversion, Consolidation and Export

Using the PoC prototype the RDF information gathered in phase 2 is converted into a tabular structure. In order to consolidate and resolve contradicting information from the different languages, Data Fusion is applied. The fusion algorithm leverages a majority voting approach for numeric values and uses a preference list for other values (English over German etc.) to filter outliers or (extraction) errors. The results are finally exported to CSV.

# Evaluation

The searches carried out during the testing phase proved that 70 of 73 selected groups could be found in DBpedia.

In contrast to the high percentage of enterprise groups matched in DBpedia, the number of retrieved attributes for the 2016 reference year is less promising. As shown in Table 1, for 70 out of 73 groups at a minimum a general description is provided.

Regarding the key figures, however, for less than 50% of the companies the number of persons employed, turnover, or total assets could be obtained by parsing DBpedia. The percentage of coverage per attribute is shown in Table 1.

Table 1. Sample Completeness Analysis of Attributes of Interest for EGR

|  |  |  |
| --- | --- | --- |
| **Attribute of the MNE** | **completeness** | **completeness in %** |
| Short Description | 70 | 95.9 |
| NACE Code | \* | \* |
| Persons employed  | 31 | 42.5 |
| Persons employed outside Europe | \* | \* |
| Total Assets | 12 | 16.4 |
| Total Assets Currency Code | 12 | 16.4 |
| Total Turnover | 27 | 37.0 |
| Total Turnover Currency Code | 27 | 37.0 |
| Country Code  | 56 | 76.7 |
| Ultimate Controlling Unit | \* | \* |
| Hyperlink to group’s Webpage | 67 | 91.8 |

The approach to determine the accuracy of the retrieved data was to manually crosscheck a sample of key figures against the officially published annual account reports by the companies on the internet.

For 31 out of 70 MNEs, the number of persons employed could be retrieved from DBpedia. Sixteen of those 31 groups were crosschecked with manually collected data.[[1]](#footnote-1)

As shown by Figure 1, the overall accuracy of the number of persons employed is very high since there is no remarkable deviation from the annual report data.



Figure 1. Persons employed scatter plot. Extracted data is shown on x-axis, manually collected data from financial reports on y-axis.

For 27 out of 73 MNEs, the turnover could be retrieved from DBpedia. Twenty-four of those 27 are crosschecked with manually collected data as shown in Figure 2.



Figure 2. Turnover scatter plot. Extracted data is shown on x-axis, manually collected data from financial reports on y-axis.

Although the overall accuracy of values is good there are few values close to zero retrieved from DBpedia. This problem occurs when a complex mixture of comma, point and currency is given.

For 12 out of 73 MNEs the total assets could be retrieved from DBpedia. Ten of those 12 are crosschecked with manually collected data.

# Conclusions

Additional sources of information such as crowdsourcing platforms, web crawling and different open data projects are seen as further opportunities to increase the quality of the EGR, its completeness and accuracy namely with aggregate indicators on the whole group level for the larger MNEs. The aim of the feasibility study was:

* to automatize at a large extend the collection of aggregated whole group figures using as input the names of the enterprise groups; and
* to evaluate the level of coverage, completeness, accuracy and timeliness.

The results from the feasibility study show that a complete automatisation was not achieved. The exported data requires further analysis and human intervention before the data is used. Regarding the coverage, the results are positive, having a high percentage of linked information based on the searching criteria. In terms of completeness, the three indicators (persons employed, turnover and assets) showed different results. The highest percentage of coverage for persons employed is still below 50% (42.5%), for turnover it is 37.0% and for assets 16.4%. The retrieved data on the three parameters showed high accuracy when compared to the figures published by the groups on their websites.

1. It is necessary to note that not every group provides an estimated or exact number of employed persons in their annual report. Only those reports serve as resource to measure the accuracy of the data parsed from DBpedia. [↑](#footnote-ref-1)