26th Meeting of the Wiesbaden Group on Business Registers

Neuchâtel, 24 – 27 September 2018

David Broska, Dimitar Nenkov (Eurostat)

Session 5 - New Data Sources

New data sources for the EuroGroups Register

1. Introduction

Eurostat governs the EuroGroups Register (EGR), the statistical business register of multinational enterprise groups (MNE groups) in the European Union. In order to create the EGR Eurostat collects relationships and enterprise group information from the national statistical business registers of the EU Member States, participating EFTA countries and from commercial data sources. The EU part of the legal units, enterprises and enterprise groups and their characteristics are therefore well-covered by the EGR.

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 the units outside of the EU and EFTA as well as on the whole group level.

Under the umbrella of Eurostat BIG DATA project, Eurostat 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 MNE groups. DBpedia is a project which extracts structured information from Wikipedia to make it publically available in a format that allows to ask sophisticated queries against Wikipedia and to link different data sets to Wikipedia data.

This paper presents the results of the feasibility study carried out by Eurostat EGR Team and Leipzig University. The objective of this prove of concept is to automatize at a large extend the collection of aggregated whole group figures using as input the names of the enterprise groups.

Currently, final checks and updates of enterprise group figures are done manually in EGR by looking into the data published in the annual accounts or websites of the enterprise groups. A population of 73 MNE groups was selected based on size and geographical diversity and provided 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 successful 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 group itself
- Timeliness - availability of data for certain reference period based on EGR cycle

2. Methodology and Interface for Data Retrieval from DBpedia

DBpedia is a project that extracts structured data from Wikipedia in order to make it publically available in a format that overcomes limitations of the latter. Wikipedia is an online encyclopaedia that is collaboratively written by volunteers. Currently about 71 000 active contributors are working on more than 47 Million articles in 299 languages.[1] 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 of the information might be quite different.

This chapter provides details about how data was retrieved from DBpedia in order to better understand the origin of the data. The technical features of DBpedia described in the following section also indicate why DBpedia is considered a more reliable source than Wikipedia to extract data on enterprise groups.

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
- Data conversion and export phase.

2.1. Phase 1: Entity Linking (Company Identification)

Every company name is sent to the DBpedia Spotlight API to identify and annotate the entity name with a DBpedia Identifier. In the default configuration, multiple language models are used ("en", "fr", "de", "nl", "it", "es", "pt", "hu"). Every language model is trained on the Wikipedia/DBpedia chapter of its corresponding language and therefore is aware of all Wikipedia pages available in its chapter. Combining these language models improves the recognition of language dependent aliases or name variations and also improves coverage for entities which have only a restricted regional importance.

Using the Spotlight API an additional filter is applied to prune the list of annotations to be of the type Organisation (dbo: Organisation [2]). One request to the API can deliver multiple candidates. Therefore every candidate has a confidence value ranging from 0 to 1.0 calculated by internal spotlight algorithms. Using the API candidates with a confidence lower than 0.5 are filtered out (for more information see the reference below[3]).

The output file gives an overview which DBpedia resource candidates were returned per language model for each enterprise name. There is also an option to adjust the number of results returned by the interface. However, it should be noted that a manual check might be necessary in order validate

[1] https://en.wikipedia.org/wiki/Wikipedia:About (2018-06-13)

[2] http://dbpedia.org/ontology/Organisation (2018-06-13)

[3] https://www.dbpedia-spotlight.org/docs/spotlight.pdf (2018-06-13)

whether the correct candidate was returned. A correct match of a DBpedia entity a provided group name should not be taken for granted.

2.2. Phase 2: Data Download

Using the Wikimedia API plus the Wikidata identifier resolved in Phase 1 the available Wikipedia article URLs for all languages are retrieved for a company. Since every chapter uses its own localized identifiers for things (e.g. Βέρνερ_φον_Ζίμενς versus Werner_von_Siemens), for every resource a normalization of identifiers is performed by using again Wikidata identifiers, in order to prepare the data fusion and consolidation in Phase 3.

Reference year information from e.g. number of employees and revenues very often is 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 editors also updated the articles of the groups for the business year 2017.

The initial strategy of the prototype is to retrieve the current information of Wikipedia. However, in order to achieve better coverage for a given reference year which is more than a year in the past Leipzig University integrated a historic mode in the prototype. To put it in a nutshell, this mode tracks all changes in the info box and merges this information into one file per reference year. Such file represents all information for one entity from a single chapter which is valid for the target reference year.

A piece of information is considered to be valid for reference year x, if the reference year x was explicitly mentioned. If multiple conflicting values for the same property and year x occur in the revision history the most recent one in history will be used to represent the company attribute for year x. If a value is typically provided with no reference year (e.g. the country or name of a company) it is only considered valid for year x, if a change to the info box occurred in year x. This is based on the assumption, that a Wikipedia editor does not change data in the info box when it is still valid.

2.3. Phase 3: Data Conversion and Export

The next step is to convert the information gathered from the info boxes into a tabular structure. Data Fusion is based on the tabular representation and starts by copying all values from the English resource (if exists), else from the German resource (if exists), else from any other language as basis for the fusion.

In the next step the financial values only valid within a reference time (assets, revenue, and employees) are cleared in the base resource and replaced by values from all languages for the target reference year. If multiple values exist for one field for a specific reference year, the one with the most occurrences will be taken to filter out outliers or (extraction) errors. For other fields the union of all values from all sources will be taken.

Based on the data derived in the two steps before multiple CSV files will be created.

3. Coverage

In order to prove the feasibility of retrieving data from DBpedia, a sample of 73 MNE groups was selected addressing groups size and geographical location diversity. These groups were taken from a data set received from commercial data source Dun & Bradstreet covering a selection of 3000 groups.

The searches carried out during the testing phase proved that 70 of those groups could be found in DBpedia. As described in more details in the following section, for 70 out of 73 at least some information could be retrieved.

4. Completeness

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 by the following table, for 70 out of 73 groups at a minimum a general description is provided. In 67 cases also a link to the company's web site is listed.

Regarding the key figures, however, for less than 50% of the companies the number of employees, turnover, or total assets could be obtained by parsing DBpedia. Table 1 presents the percentage of completeness per attribute.

Table 1 - Completeness of DBpedia per group attribute of 73 MNE groups

Attribute	Completeness	Completeness in %
Group description	70	95.9
Group main activity code	0	0.0
Group persons employed	31	42.5
Group persons employed outside the EU	0	0.0
Group total assets	12	16.4
Group total assets currency	12	16.4
Group turnover	27	37.0
Group turnover currency	27	37.0
Country of group (global decision centre)	56	76.7
Group web address	67	91.8

5. Accuracy

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

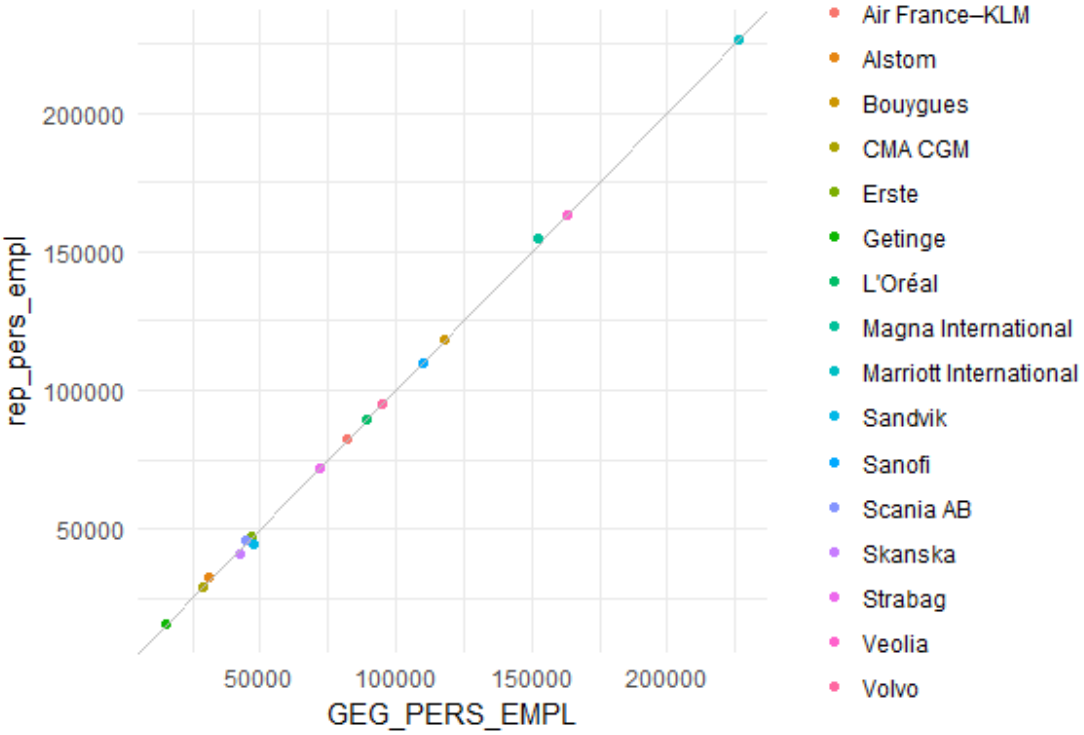
The following three figures visualize the differences between the automatically parsed data from DBpedia (GEG) and manually retrieved data from annual reports of multinational companies (rep) on reference year 2016. If there is no difference between these two values, they should align with the 45° line since $y = x$.

5.1. Persons Employed

For 31 out of 70 MNEs the number of employees could be retrieved from DBpedia. 16 of those 31 were cross checked with manually collected data.[4]

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

Figure 1 - Accuracy of number of employees data from DBpedia (GEG) and from annual reports (rep)

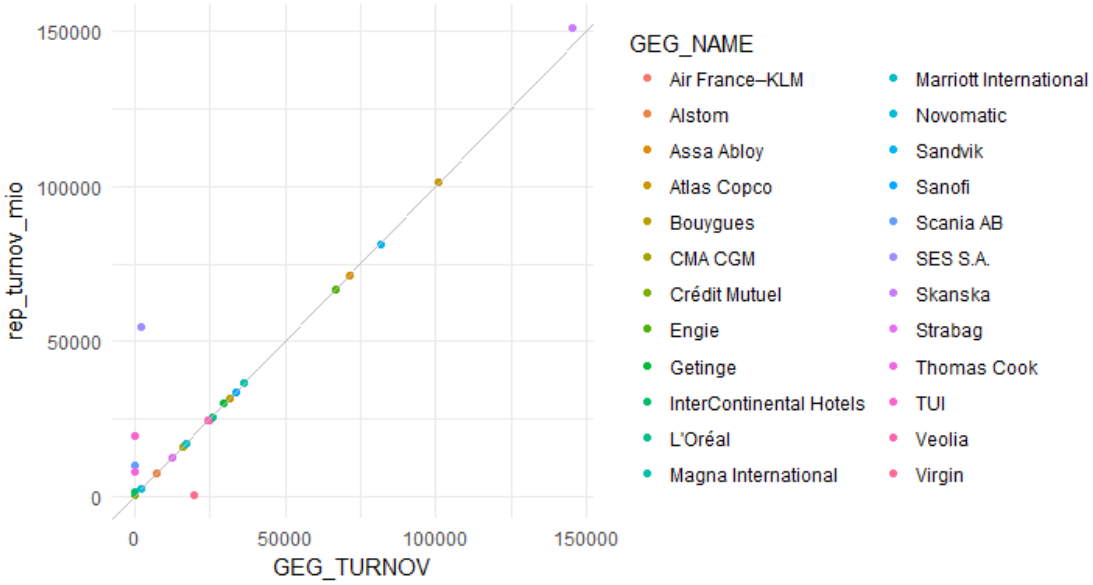


5.2. Turnover

For 27 out of 73 MNEs the turnover could be retrieved from DBpedia. 24 of those 27 were cross checked with manually collected data. Figure 2 presents the accuracy of turnover data from DBpedia compared to data from annual reports.

Figure 2 - Accuracy of turnover data from DBpedia (GEG) and from annual reports (rep)

[4] It is necessary to note that not every group provides an estimated or exact number of employees in their annual report. Only those reports serve as resource to measure the accuracy of the data parsed from DBpedia.



Although the overall accuracy of values is remarkable, there are few values close to 0 retrieved from DBpedia. This problem occurs when a complex mixture of comma, point and currency is given. Table 2 illustrates such cases. In fact, the modifier (million, billion etc.) was not interpreted correctly. There might be the possibility to optimize parsing numbers from text, but it seems that there will be no 100% correct extractions for every single case in DBpedia.

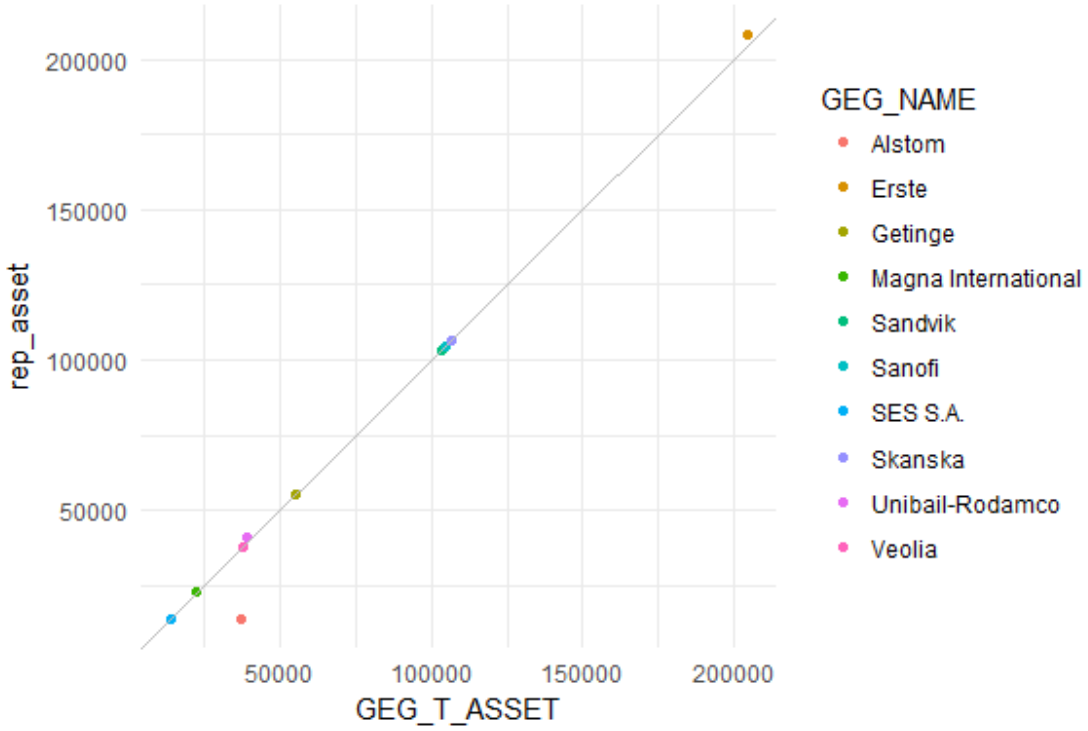
Table 2 - Examples of problematic turnover figures as retrieved from DBpedia

Group name	Annual report value (rep_turnov_mio)	DBpedia value (GEG_TURNOV)
Crédit Mutuel	332	0.001465
InterContinental Hotels Group	1715	0.001715
Thomas Cook Group	7812	0.007812
TUI Group	19584	0.017184
Scania AB	10184	0.103927

5.3. Total assets

For 12 out of 73 MNEs the total assets could be retrieved from DBpedia. 10 of those 12 are cross checked with manually collected data. Figure 3 presents the accuracy of the total assets data from DBpedia compared to data from annual reports.

Figure 3 - Accuracy of total assets data from DBpedia (GEG) and from annual reports (rep)



The outlier in total assets for Alstom should be regarded as a reminder that numbers from DBpedia comes from different Wikipedia pages and therefore can be reported in different currencies. The total assets are of 13622 Mio. EUR in the annual report referenced here whereas DBpedia reports 37160 Mio. USD. (Still, these values differ considerably after converting currencies by the exchange rate as of December 2016).

6. EGR cycle compared to timeliness of available data in DBpedia

As described in section 2.2. Phase 2: Data Download the interface includes a historical mode that enables the user to retrieve data on enterprise groups even if the Wikipedia has already been updated with new data. For example, the data used in this report for the 2016 reference year could be retrieved although this information is no longer available in the current Wikipedia chapter.

Due to the delay with which the EGR provides data on enterprise groups this feature is essential. Consider, for example, that in 2018 the financial statements of the enterprise groups for 2016 are finalized by EGR. The info box may have already been updated with the financial statements of the 2017 reference year. Despite this fact the interface for data retrieval offers the possibility to lookup data for the 2016.

Table 3 shows the number of retrieved values for the turnover, total assets and the number of employees. Since the data is updated to Wikipedia at the earliest with the release of the annual report of the enterprise groups, information for 2017 is still missing. However, 58 values for the 2015 reference year and 70 values for the 2016 reference year were uploaded to Wikipedia until October 2017 (the date of the DBpedia database snapshot).

Table 3 - Number of retrieved values for turnover, assets and employees from DBpedia

Reference year	Turnover	Assets	Employees
2014	11	0	17
2015	22	8	28

2016	27	12	31
2017	0	0	3

7. 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 the units outside of the EU and EFTA as well as the aggregate indicators on the whole group level.

The aim of Eurostat is to carry out feasibility studies in this context. With the current prove of concept it was tested to automatize at a large extend the collection of aggregated whole group figures using as input the names of the enterprise groups and evaluate the level of coverage, completeness, accuracy and timeliness.

The results from the feasibility study show that a complete automatization was not achieved. The exported data would require 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 3 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.