Web Intelligence on multinational enterprise groups data

UNECE Group of Expert on Business Registers

26 – 29th September 2022

Session 2: 'The use of administrative data, big data, and other data sources'

Smart data for MNE groups

- ✓ Web Intelligence for statistical production aims to develop statistics gathering information from the Internet, using innovative data collection methods.
- ✓ This study was started in order to retrieve information about enterprise groups operating in the EU and EFTA countries, using public sources on the web.
- ✓ In this presentation a Web Intelligence process is defined to assess quality level of usable sources.

Smart data for MNE groups

- ✓ The Study was based on a selected sample of 200 MNE groups
- ✓ Objectives of the Web intelligence process:
  1. Discovering phase: detect the possible relevant data sources to produce information on groups.
  2. Implementation phase: assessment of the quality of data sources based on EGR data and build the database.

Discovering phase

✓ Most relevant identified data sources:

Source	Available information	Structured information	Information level
Annual Reports	Almost 100%	x	Group
GLEIF	About 90%	✓	Group/Legal Unit
PermID	About 90%	✓	Group/Legal Unit
Wikidata	About 88%	✓	Group
Wikipedia	About 80%	✓	Group
DBPedia	About 60%	✓	Group/Legal Unit
Open Corporates	About 25%	✓	Group/Legal Unit
EDGAR	About 20%	✓	Group

Discovering phase

- ✓ Group level: information on variables vs sources

Variable	Annual Report	GLEIF	PermID	Wikidata	Wikipedia	DBPedia	OpenCorp	EDGAR
Name	✓	✓	✓	✓	✓	✓	✓	✓
Website	✓	x	x	✓	✓	✓	x	✓
Number of person employed	✓	x	x	✓	✓	✓	x	✓
Asset	✓	x	x	✓	✓	✓	x	✓
NACE	✓	x	x	✓	✓	✓	x	✓
Turnover	✓	x	x	x	x	x	x	✓
Year	✓	x	x	✓	x	x	x	✓

Discovering phase

- ✓ Legal Unit level: information on variables vs sources

Variable	Annual Report	GLEIF	PermID	Wikidata	Wikipedia	DBPedia	OpenCorp	EDGAR
Name	✓	✓	✓	✓	x	✓	✓	✓
Website	✓	x	✓	✓	x	✓	x	✓
Number of person employed	✓	x	x	✓	x	✓	x	✓
LEI number	x	✓	✓	✓	x	x	x	x
Address details	✓	✓	x	x	x	x	x	✓
City	✓	✓	x	x	x	✓	x	x
Country code	x	✓	x	x	x	✓	x	✓
Year	✓	x	x	x	x	x	x	✓
Primary National ID	x	✓	x	x	x	x	x	✓
Activity status	x	✓	✓	x	x	x	x	x
Date of incorporation	x	✓	x	x	x	x	x	x

Discovering phase

- ✓ Qualitative data is widely available: well-structured information comes from GLEIF and PermID.
- ✓ Quantitative data are less available: Wikipedia, Wikidata and EDGAR provide reference dates and some time series for specific variables.
- ✓ Information on MNE Group subsidiaries is available in GLEIF and Wikidata
- ✓ The Annual reports and EDGAR contain almost all information, but were not considered in the study due to the higher level of complexity for automatic information extraction.

Implementation phase

Assessment of the quality of data sources based on EGR data: identify the coverage and the comparison of each variable

Process steps:

1. Link the MNE groups used for the Study with the those present in the EGR
2. Check continuity of Group ID between 2019 and 2020 RY
3. Variable mapping between the used sources and the EGR variables
4. Define ad-hoc procedures for each analyzed variable to obtain the highest level of information from public sources.

Implementation phase

Assessment of the quality of data sources based on EGR data: identify the coverage and the comparison of each variable

Analysed variables at MNE group level:

- ✓ UCI-CC: Ultimate Controlling Institutional unit Country Code
- ✓ Employment: the number of persons employed
- ✓ Turnover: totals invoiced of the MNE group
- ✓ Assets: comprise total economic resources
- ✓ Number of legal units: as part of a MNE group

MNE group ID continuity

EGR 2019 vs EGR 2020:

- ✓ The initial list of MNE groups was based on EGR 2019 frame.
- ✓ In 5% of the cases, a change in group name or country code was identified:
  - in 2.5% of them, this happened without group identifier changes,
  - while in the other 2.5% of cases, a new group identifier was required.

UCI Country Code comparison

- ✓ The information on the country code of the ultimate controlling institutional unit was present in all the public sources used

Source	TOT Linked	Same CC	Different CC	*Priority*
GLEIF	*98%*	*91%*	*9%*	*1*
WIKIDATA	*82%*	*90%*	*10%*	*2*
PERMID	*98%*	*89%*	*11%*	*3*
WIKIPEDIA	*56%*	*88%*	*11%*	*4*
OPENCORP	*11%*	*90%*	*11%*	*5*
ALL SOURCES	*99%*	***90%***	***10%***	

- ✓ Integrated sources percentage of similarity *90%*

Numerical variable comparison methodology

- ✓ We first defined the relative difference of each variable for the same MNE group as:

$$RDVar = \frac{variable\ value\ from\ public\ source - variable\ value\ from\ EGR}{variable\ value\ from\ EGR}$$

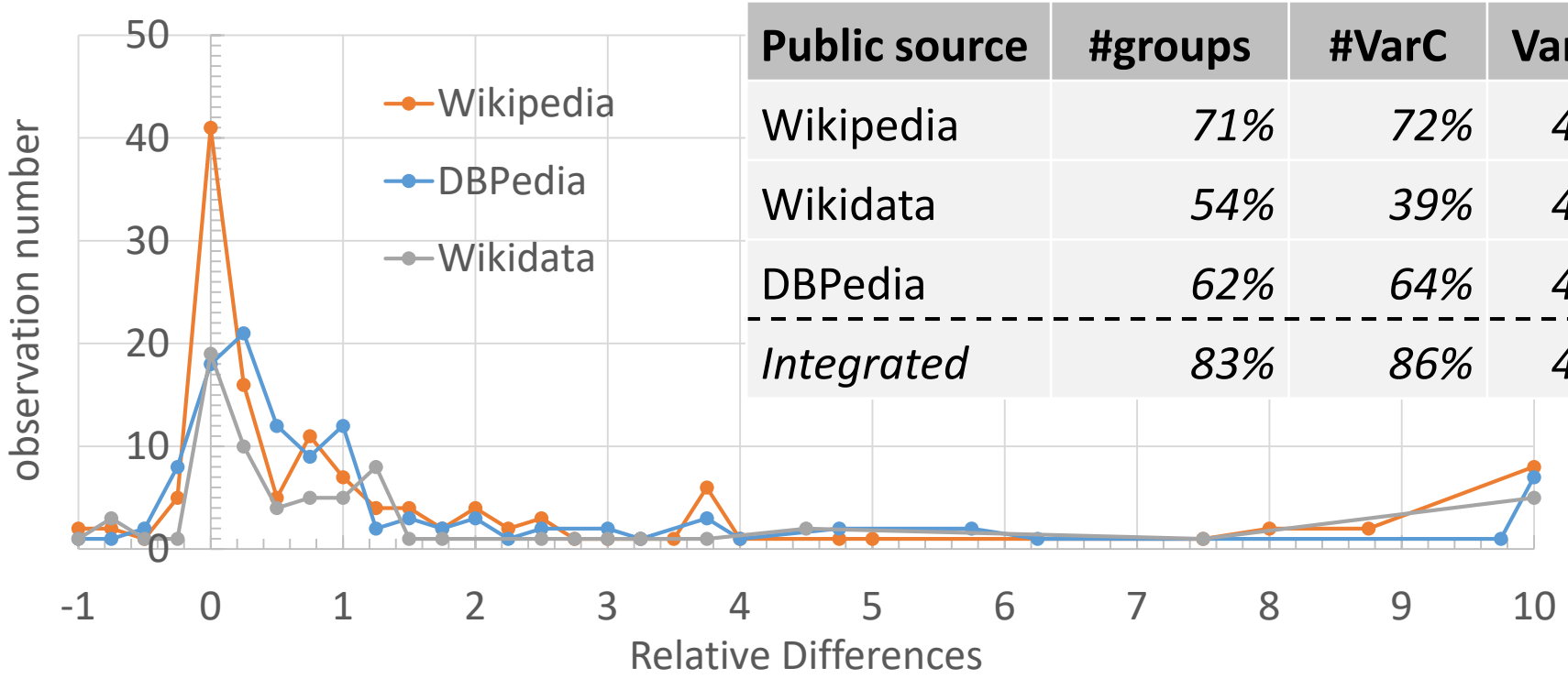
- ✓ Therefore, we defined the variable quality indicator as:

$$VarQI = \frac{\{\#VarC : |RDVar| \leq 0.5\}}{\#VarC}$$

- ✓ where $\#VarC$ is the number of comparable observations for each variable, i.e. this is linked to the number of public observations actually comparable with the EGR values.

Comparison of Number of person employed

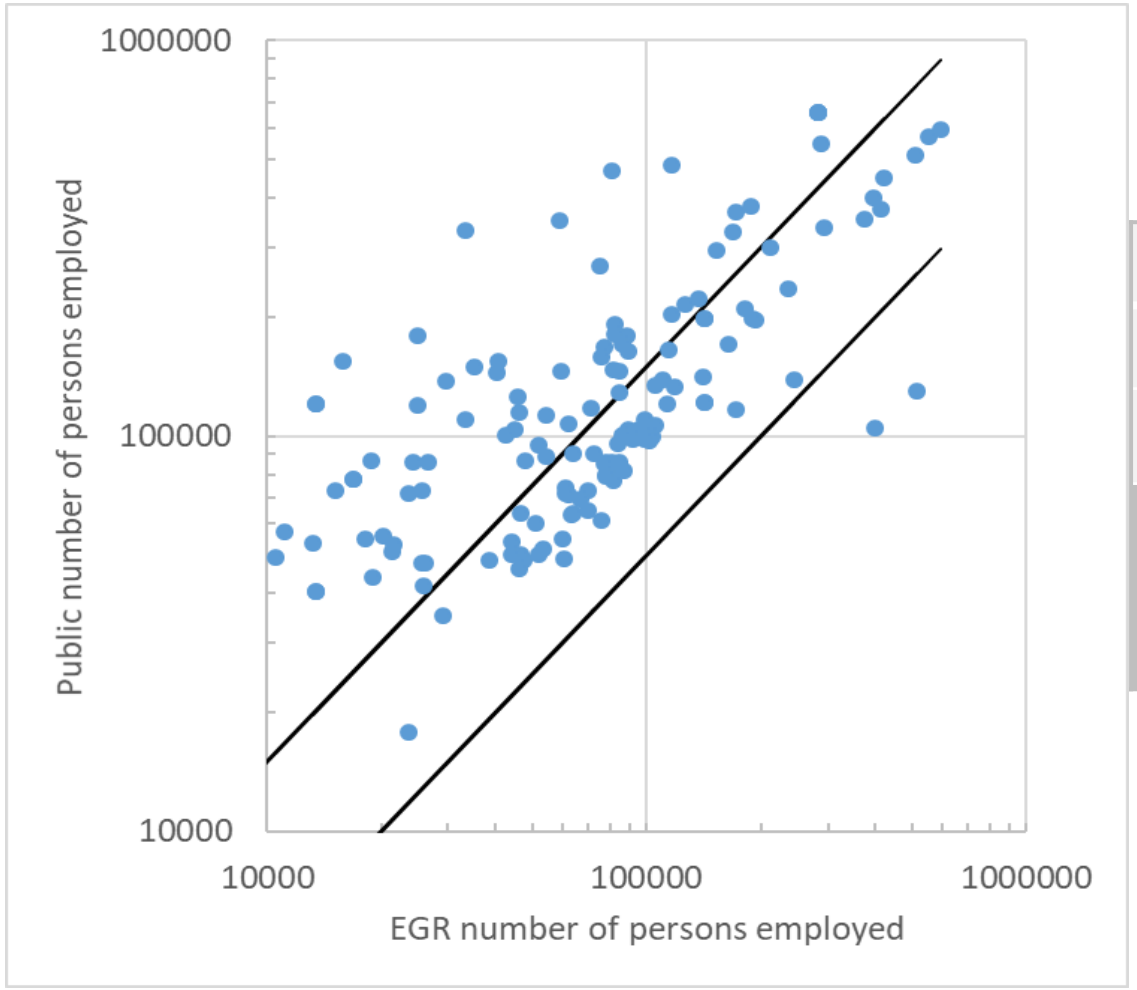
✓ Number of observations versus variable relative differences:



Public source	#groups	#VarC	VarQI	Priority
Wikipedia	*71%*	*72%*	*47%*	*1*
Wikidata	*54%*	*39%*	*43%*	*2*
DBPedia	*62%*	*64%*	*42%*	*3*
Integrated	*83%*	*86%*	*43%*	*-*

Comparison of Number of person employed

✓ Integrated view – scatter plot: Employment quality indicator = 43%



Integrated *EMPLOYED*			
		EGR	
		NULL	**NOT NULL**
PUBLIC	**NULL**	*1%*	*15%*
	NOT NULL	*1%*	*82%*

Comparison of Turnover and Asset

✓ Information retrieved from public sources:

TURNOVER

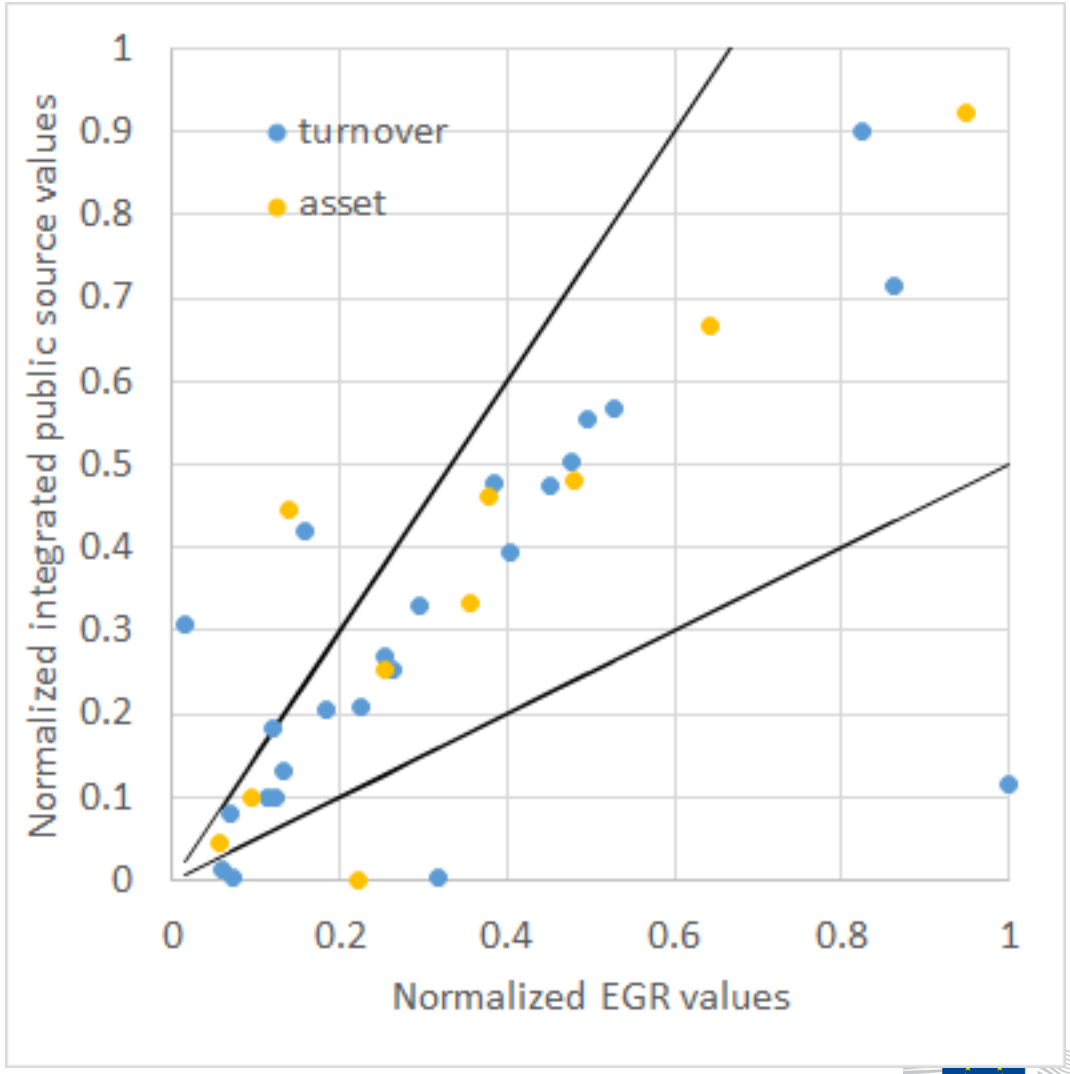
Public source	#groups	#VarC	VarQI	Priority			
Dbpedia	*46%*	*16%*	*79%*	*1*			EGR
Wikidata	*73%*	*18%*	*67%*	*2*			NULL
Integrated	*78%*	*20%*	*69%*	*-*	*Public values*	*NOT NULL*	*65%*

ASSET

Dbpedia	*41%*	*6%*	*80%*	*1*			EGR
Wikidata	*49%*	*13%*	*75%*	*2*			NULL
Integrated	*60%*	*13%*	*82%*	*-*	*Public values*	*NOT NULL*	*53%*

Comparison of Turnover and Asset

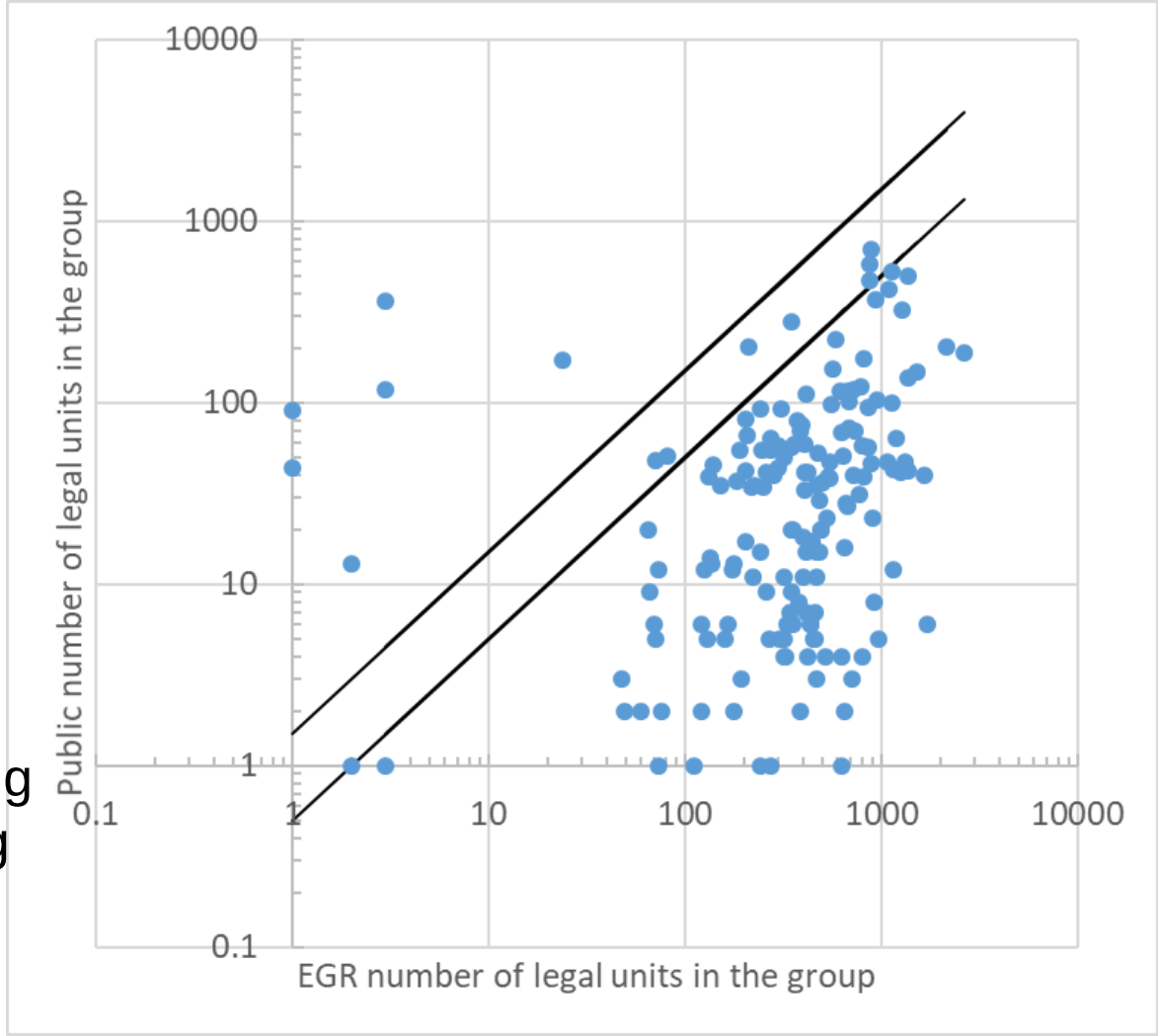
- ✓ Integrated view – scatter plot:
  - *TurnoverQI= 69%*
  - *AssetQI= 82%*



Comparison of Number of LEUs

Public source	#groups	VarQI	Priority
GLEIF	*93%*	*7%*	*1*
Wikidata	*80%*	*3%*	*2*
Integrated	*93%*	*7%*	*-*

- ✓ Integrated view – scatter plot:
  - NumLUnitQI= 7%
- ✓ GLEIF can improve the EGR by completing the list of units owned by groups operating in Europe to a level of 0.2%.



Conclusion

- ✓ Public sources appear to be interesting to fill in any missing values and to assess and consolidated information of the MNE groups in the EGR
- ✓ The gain on MNE groups seems to be positive for the country of the ultimate controlling unit (UCI), turnover and assets.
- ✓ Regarding the information on employment and legal unit, further analysis is needed to understand the discrepancies found.
- ✓ Public sources can also be used as early detection of changes to trigger necessary actions to maintain the EGR.

Keep in touch

Antonio.LAURETI-PALMA@ec.europa.eu

Alexandre.DEPIRE@ec.europa.eu

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

European Commission

Thank you

European Commission