UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Expert meeting on Statistical Data Collection

(12 -14 June 2023)
https://statswiki.unece.org/x/MADUE

Data collection methods to produce new enterprise variables using new data sources

Scalfati F., Bianchi G., Salamone S.

ISTAT (Italy)

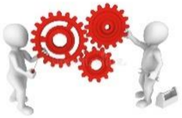
Outline

- ✓ Objective
- ✓ Data Collection strategy
- ✓ Data Collection Process
- ✓ Experimental result
- ✓ Conclusion

Objective

The **aim** of this work is to produce a **statistical framework** able to extract detailed **information** on the **innovative capacity of enterprises** and **produce new statistical variables**, by means a **data analytics approach**.

Data Collection strategy (1/2)



This work **combines multiple sources** (big data, survey data and registers) in order to produce **indicators** that provide the **profile of the enterprises**. In particular, the identification of the patenting enterprises allows linking them to the structural characteristics and provides additional dimensions available for this goal.

The data source used, is the most complete and updated **database on patents published by the European Patent Office** (EPO) which it acquires data from the EPO's master bibliographic database. The target data in EPO are the applicants based in Italy published patent/s.

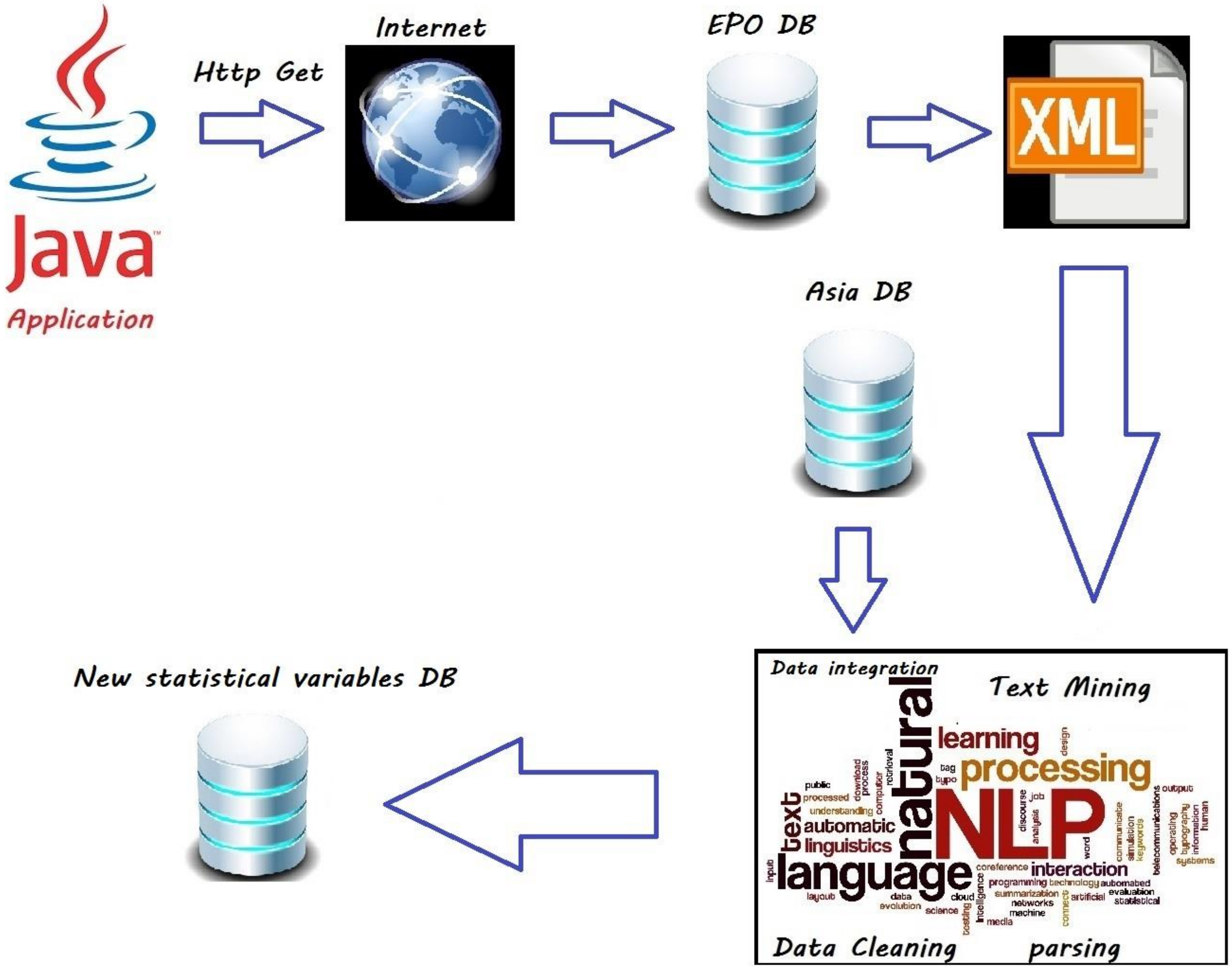
Data Collection strategy (2/2)

The planned statistical output has as reference population the active **enterprises available** from the *Italian National Business Register (**ASIA**)*.

The proposed approach for **collecting statistical information** on the innovative capacity of enterprises **acquires European patent publications** in text format **using APIs and web scraping techniques**.

It **integrates** the **extracted information** with **statistical registers** and **surveys** and produces **new statistical output** by using **text mining** and **machine-learning techniques**.

Data Collection Process



Data characteristics

The procedure **collects** the following macro variables:

- Name of the applicant, owner and inventor
- Localization information on the residence of the three subjects
- Type of patent
- Date of publication of the patents
- Patent filing date
- IPC code (International Patent Classification)

All data collected refers to the **geographic origin of the applicant/owner (**country of residence).

Data integration

- Integration step is based on **record linkage** procedure to match **micro-data** on patent application from the **EPO** server with the data available from the Italian Official **Business Register** (ASIA).

- Availability of data on an annual basis is preliminary to allow the subsequent integration phase.

- For the match between the two sources it is necessary to know the **year of publication** of the patent to identify whether the company was **active** in the reference year.

- Data collection procedure must to extract complete information, **without duplicates** in order to allow unambiguously identification.

Experimental results

- In the case study **8000 URLs** have been extracted from **EPO DB**.

- The procedure acquired **the related patents** from **the European Publication server**.

- Each record is composed of about **40 variables:** proponent (applicant, owner, inventor), personal data, type of patent, patent features, references, claims

- Data refers **to Italian patents**

- Some **output indicators**: rate of proponent, rate of patents, territorial distribution, thematic distribution

Conclusions

The **innovative capacity** of **enterprises** and **institutions**, can be **filled** with **indicators** that provide the **profile** of the **enterprises**.

The patenting enterprises allow to **produce new information** by **linking** with **structural** and **economic characteristics**.

Patent statistics are effective **proxies** for **measuring** and **monitoring innovative** activities spread across a **territory**.

This **automatic approach reduce** the **burden** on enterprises.

It's a **difficult task** because **extracts text** from website and **uses text mining** and **machine learning techniques** to produce **new statistical variables** in **reasonable time.**

Contacts:

Scalfati Francesco (scalfati@istat.it)
Bianchi Gianpiero (gianbia@istat.it)
Sergio Salamone (sesalamo@istat.it)