Web Mining to integrate and validate information from the statistical register of active enterprises (ASIA)

| MONICA CONSALVI - ISTAT | BARBARA GENTILI - ISTAT | FLAVIO PANCELLA - ISTAT | CATERINA VIVIANO - ISTAT | GIANPIERO BIANCHI - ISTAT |

FRANCESCO SCALFATI - ISTAT | DONATO SUMMA - ISTAT | MASSIMILIANO AMARONE - ISTAT | DOMENICO APRILE – ISTAT |

ENTERPRISES CONSIDERED IN THE EXPERIMENTATION

- The design and implementation phase was carried out on a **sample** of about 100,000 enterprises. The unit sample was extracted from the latest available Asia-business register release, stratified by size in terms of employment / business volume, sector of economic activity and legal form.

ADOPTED TECHNIQUES

- *Web Scraping* techniques for web data acquisition: **i)** scraping from a specific URL; **ii)** scraping through the use of search engines in batch mode; **iii)** specialized scraping from thematic portals.
- *Text Mining* techniques, with the use of *Natural Language Processing*, for the analysis and extraction of information (data, documents, metadata, files, etc.) within natural language texts.

ENTERPRISE IDENTIFICATION ON THE WEB

- In case of *URL from administrative source*, a syntactic validation of the string, a check of the recurring errors and a verification of the authority (domain) of the address were applied.

- In case of *missing URL*, information was downloaded directly from thematic business portals or URL retrieval techniques were applied (by executing batch queries on search engines using company master data) combined with machine learning techniques to estimate the probability of the identified URL.

URLs from ASIA business register	63 %
URLs extracted from the WEB	17 %
URLs extracted from certified mail addresses	15 %
URLs from thematic portals	5 %

In both approaches it is necessary to proceed with the correct identification of web addresses that certifies an exact correspondence between the websites and the enterprises contained in the register:

✓ Extraction of the main information on the enterprise's website (Fiscal code, VAT code, Company name, Address…) by using *Information Retrieval* techniques through *pattern matching* on strings.

✓ Comparison with the information available in the ASIA business register through matching techniques and *string similarity metrics* (Jaro-Winkler, Levenshtein, etc.)

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: DESCRIPTION OF THE ACTIVITY

Extraction of a short description of the enterprise activity

The short **description** of the **enterprise activity** has been acquired through:

✓ **Meta-information (meta- tag) on web pages**
✓ **The result of the batch querying of Bing search engine**

Unit code	Enterprise name	Activity description
xx1	AAA SRL	Company specialized in calendering and bending of pipes and profiles..
xx2	BBB Security	Installation and maintenance of security alarm systems, telecommunications and home automation.
xx3	DDD SPA	Shop of nautical accessories.
xx4	EEE SNC	Fabrics of value to satisfy every customer's need.
xx5	FFF SPA	Production of flat rolled steel, carbon steel welded tubes, stainless steel welded tubes, stainless steel precision belts.

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: SHARE CAPITAL

Extraction of the value in euros of the Share Capital contained in the enterprise's website.

✓ The value of the share capital has been acquired through the use of Information Retrieval techniques
✓ The share capital extracted was related to the 14% of the companies (refers to those that expose the registered capital on the website)

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: SOCIAL MEDIA PRESENCE

Identification of the enterprises that show *Social Media links* on their websites **(Facebook, Twitter, Instagram ecc.):** presence of links to an account or profile of the respondent enterprise on a specific social media

1.	Facebook	62.545
2.	Twitter	32.698
3.	Youtube	19.007
4.	Instagram	12.439
5.	Pinterest	2.070
6.	Thumblr	1.415
7.	Flickr	1.411
8.	Xing	1.057
9.	MySpace	304
10.	Slide Share	234
11.	Picasa	112

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: TELEPHONE NUMBER

Extraction of the telephone number contained in the enterprise website

Distribution by type of telephone number:

Enterprises with landline phone	57 %
Enterprises with fax	39 %
Enterprises with mobile phone	24 %

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: ADDRESS

Extraction of the street addresses contained in the enterprise website

✓ Extraction of the street address contained in the enterprise website through the use of *pattern matching* techniques

✓ A total of **100.000 addresses** have been divided into **administrative offices** and **local units**

Administrative offices addresses	70 %
Local units addresses	30 %

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: E-MAIL ADDRESS

Extraction of the e-mail address contained in the enterprise website

✓ The enterprise email address has been extracted from the enterprise's website using *pattern matching* techniques

✓ Distribution by type of e-mail address:

Enterprises with e-mail address	90 %
Enterprises with certified e-mail address	5 %

ADDITIONAL INFORMATION OBTAINED FROM THE WEB: PRESENCE OF ENTERPRISE BALANCE SHEET OR CATALOGUES IN PDF FILE FORMAT

Identification and download of pdf files on the enterprise website related to financial statements, product catalogs and other categories

✓ **Identification of PDF** files related to financial statements, product catalogs and other categories by analyzing the content of the files

✓ **Download** of the identified files

✓ **Archiving** by category