Web Intelligence Hub

Use and analysis of web scraped data for different statistical domains

Data science for better decisions
Brussels, 17th December 2019

Martina Hahn
European Commission (Eurostat)
Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics) - 2018

- Implement practical and mature cases (prototypes)
- Review and adapt
  - Business architecture, ~ processes
  - IT infrastructures
  - Methodological and quality frameworks
  - Governance structures
- Legal framework
- Architectural elements
  - Privacy by design
- Dedicated smart statistics communities
- Standards
- Skills
- Partnerships
- International cooperation
Principles of Trusted Smart Statistics

- Multi-source statistics
- Multi-purpose data sources
- Layered organisation: the hourglass model
- Modular methodological frameworks
- Pushing computation out
- Use data without sharing
Trusted Smart Statistics Centre (TSSC) 
Web Intelligence Hub (WIH)

A bundle of capabilities to support the collection, processing, reuse and analysis of web data resource (web pages, APIs ...) for producing statistics

- Online job vacancies advertisement
  - Skills, job vacancies
- Enterprise websites
  - Business registers, jobs, information society
- Wikipedia / EDGAR / ESEF
  - EuroGroups Register

Looking for a job?
Web Intelligence Hub - Services

- Provide support to ESS partners in:
  - Data acquisition (web scrapping, APIs)
  - Trans-national data agreements
  - Partnership models for national data agreements
  - IT infrastructure and tools
  - Analytical services (e.g. NLP)
  - Methodology
  - Regulatory aspects
  - Skills (training material)
  - R&D collaboration
  - Governance
Web Intelligence Hub - Principles

Some principles
- ESS hub - Serving European and national needs
- Modular structure
- Defined processes and products to be guaranteed
- Priority to working together, possibility to act individually
- Transparency as much as possible
  - Common used processes should be certified and audible
  - Lineage of data and processes
- Intermediate products usable by all partners
Web Intelligence Hub – Use cases

- Initial use cases
  - Online Job Vacancies
  - New Data Sources EGR

- Potential future use cases
  - Tourism accommodations
  - New forms of labour (Gig economy)
  - Online prices
  - Information society, e.g. E-Commerce
  - Economic nowcasting
WIH use case: Online Job Advertisements (OJA)
Online job advertisements

- Advertisements published on the World Wide Web on job vacancies available in enterprises
- Include data on
  - characteristics of job (occupation, location, ...)
  - characteristics of employer (economic activity, ...)
  - requirements (education and skills, ...)
- Partly available only as natural language data
  - requires specific methodologies of processing and analysis but also much higher information richness and avoids pre-conceived classifications (e.g. important to identify emerging skills)
The objects of interest

- Not published
- "Ghost vacancy"
- Published
- Published online
- Published offline

Vacant post
Extracting data from an OJA

Web scraping: extracting data from web pages

DATA SCIENTIST
Location: Luxembourg
Application deadline: Monday, 30 September 2019
Reference number: 100

Description
This role will directly support the Head of Data Labs in the design and implementation of advanced analytics as part of the Group-wide Data and Analytics strategy. We are seeking a high calibre creative thinker who can turn business problems into solutions by applying sophisticated and targeted analytics. The role will be project based, working as part of the wider Data Labs team to put analytics into practice across the organisation.

Title: Data scientist
Area: Luxembourg
Time: Saturday, 30 September 2019

Structured data
Natural language data
Data pipeline

Ingestion

Pre-Processing

Information Extraction

ETL

Presentation Area
Extracting data from an OJA natural language
# Extracting data from an OJA (Cedefop)

<table>
<thead>
<tr>
<th>Variable</th>
<th>From structured fields</th>
<th>From natural language</th>
<th>Overall precision rate (natural language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation</td>
<td>4.74%</td>
<td>100.00%</td>
<td>85%</td>
</tr>
<tr>
<td>Education level</td>
<td>4.50%</td>
<td>99.86%</td>
<td>99%</td>
</tr>
<tr>
<td>Experience</td>
<td>2.68%</td>
<td>45.64%</td>
<td>n.a.</td>
</tr>
<tr>
<td>Contract</td>
<td>29.76%</td>
<td>79.64%</td>
<td>99%</td>
</tr>
<tr>
<td>Economic activity</td>
<td>39.92%</td>
<td>100.00%</td>
<td>92%</td>
</tr>
<tr>
<td>Working hours</td>
<td>24.14%</td>
<td>64.66%</td>
<td>99%</td>
</tr>
<tr>
<td>Place</td>
<td>88.42%</td>
<td>97.86%</td>
<td>n.a.</td>
</tr>
<tr>
<td>Salary</td>
<td>23.32%</td>
<td>21.38%</td>
<td>n.a.</td>
</tr>
<tr>
<td>Skill</td>
<td>0.81%</td>
<td>90.43%</td>
<td>n.a.</td>
</tr>
<tr>
<td>Release date</td>
<td>100.00%</td>
<td>100.00%</td>
<td>n.a.</td>
</tr>
<tr>
<td>Expire date</td>
<td>13.18%</td>
<td>100.00%</td>
<td>n.a.</td>
</tr>
</tbody>
</table>
Main skills required for Office professionals

- Sales activities
- Team building
- Project management
- Use of Microsoft Office
- Economics
- Accounting
- Think creatively
- Online analytical processing
- Plan digital marketing
- Use a computer
- Tolerate stress
- Sales strategies
- Adapt to change
- Customer segmentation
- Use communication techniques
- English
- Customer service
- Foreign languages for international careers
- Work as a team
- Underwrite vehicles
- Create solutions to problems
- Software
- Social media marketing
- Management in a sales department
- Retail
- Customer relationship management
- Industrial design
- Sales argumentation
- Department processes
- Marketing
- Sourcing techniques
- Customer service
- Communication
- Teamwork principles
- Sales
- Adapt to changing situations
- Assist customers
- Project leadership
- Human resource management
- Persuasion
- Media relations
- Use economic tools
- Teamwork
- Production
- Convert
- Productivity
- Bank
- Area sales manager
- Industrial design
- Business development
- Planning
- Business plan
- Business development
- Sales
- Planning
- Business plan
- Business development
- Sales
Occupations for which the skill “Use Microsoft office” is most frequently required

Skills (ESCO version 1)
(mentioned in at least 5% of vacancies)

- Adapt to change
- Work as a team
- Use a computer
- English
- Teamwork/Principles
- Assist customers
- Use Microsoft office
- Foreign languages for intern...
- Create solutions to problems
- Adapt to changing situations
- Communication
- Think proactively
- Use office systems
- Customer service
- Show responsibility
- Office software
- Project management
- Work independently
- Team building
- Develop strategy to solve pr...
- Tolerate stress
- Adjust priorities
- Think creatively
- Work in teams
- Prioritise tasks
- Manage time
- Analyse problems for opport...
- Demonstrate enthusiasm
- Lead a team
- Quality standards
- Provide leadership
Enhanced job vacancies statistics

Figure 5: Nowcasts based on the S-ARIMA-X time series model.

Figure 6: Number of job vacancies as a proportion of working age population.
Web Intelligence Hub – Main milestones

- Launching phase: 2019
  - Agreements with Cedefop for OJA: Dec 2020
- Project phase: 2020 – 2023
  - WIH base system 1st live release: Dec 2020
  - Use case OJA: Dec 2020
  - Use case MNE: Dec 2021
  - WIH base system final release: Dec 2023
- Process phase: 2024 – Steady state
  - Additional use cases
- Steady state
  - Around 6 – 8 use cases running