



# ***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

## ESS Governance

### Web Intelligence Hub

Online  
job  
vacancies

New Data  
Sources  
EGR

...

Community  
of experts

Community  
of experts

Community  
of experts

### Trusted Smart Surveys Hub

Time  
Use  
Survey

House-  
hold  
Budget  
Survey

...

Community  
of experts

Community  
of experts

Community  
of experts

### Mobile Network Operator Data Hub

Demo-  
graphic  
statistics

Regional  
statistics

...

Community  
of experts

Community  
of experts

Community  
of experts

...

...

...



European  
Commission

# 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



Matching Skills Demand (CEDEFOP JVs) and Supply (EURES CVs):  
Absolute and relative frequencies



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



# 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 accomodations***
  - ***New forms of labour (Gig economy)***
  - ***Online prices***
  - ***Information society, e.g. E-Commerce***
  - ***Economic nowcasting***

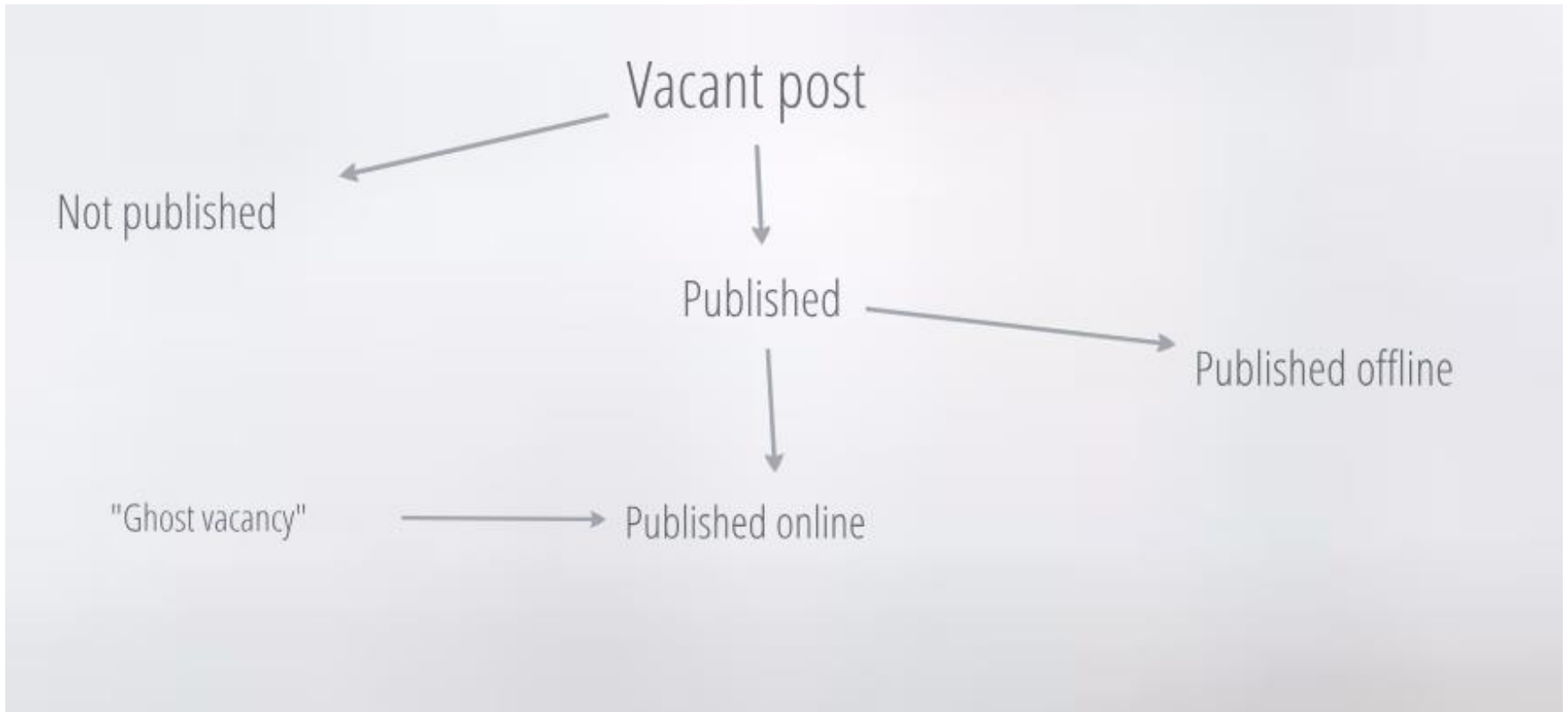


# **WIH use case: Online Job Advertisements (OJA)**

# Online job advertisements

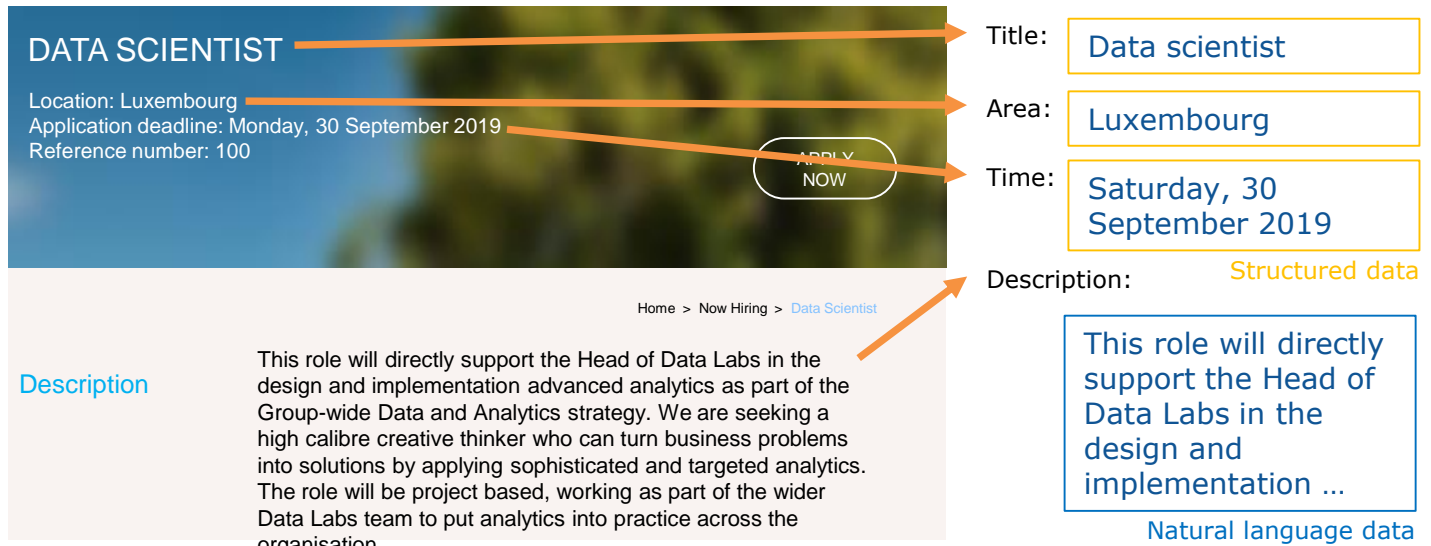
- Advertisements published on the World Wide Web on job vacancies available in enterprises
- Include data on
  - **characteristics of job (occupation, location, ...)**
  - **characteristics of employer (economic activity, ...)**
  - **requirements (education and skills, ...)**
- Partly available only 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



# Extracting data from an OJA

## Web scraping: extracting data from web pages



The image shows a screenshot of a job advertisement for a Data Scientist. The advertisement is divided into two main sections: a header section with key details and a description section. Orange arrows point from specific text in the header to corresponding fields in a structured data table on the right. Another arrow points from the description text to a field in a natural language data table on the right.

|              |                             |
|--------------|-----------------------------|
| Title:       | Data scientist              |
| Area:        | Luxembourg                  |
| Time:        | Saturday, 30 September 2019 |
| Description: | Structured data             |

**Description**

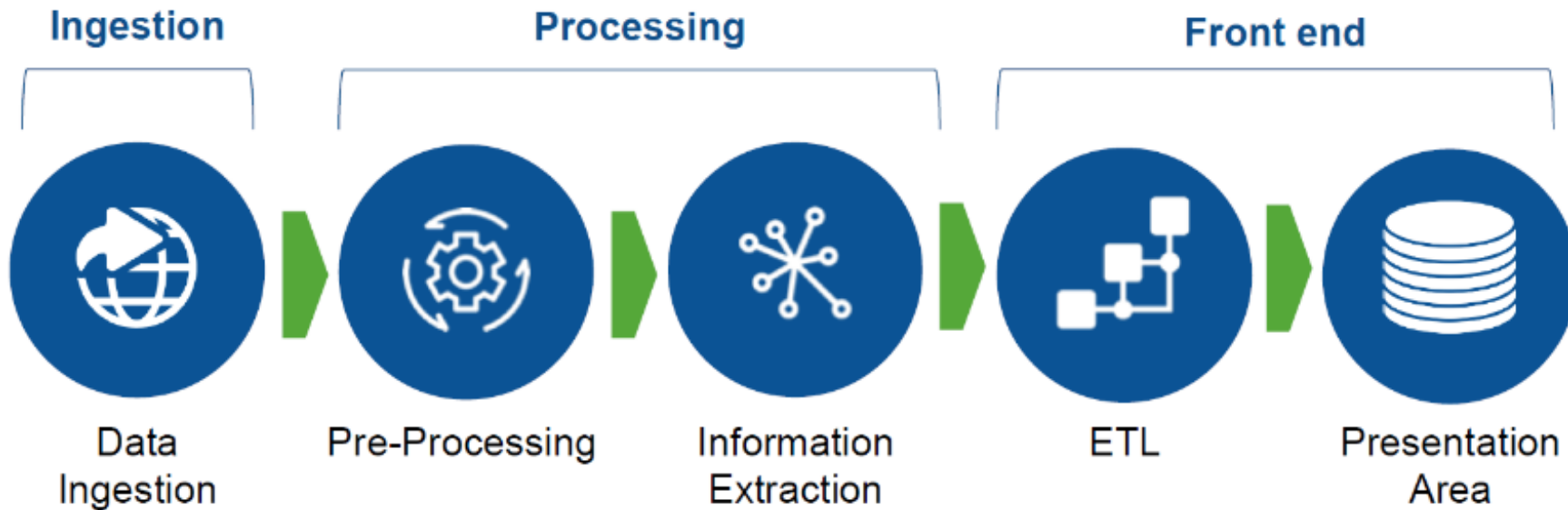
This role will directly support the Head of Data Labs in the design and implementation advanced analytics as part of the Group-wide Data and Analytics strategy. We are seeking a high calibre creative thinker who can turn business problems into solutions by applying sophisticated and targeted analytics.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

**Natural language data**

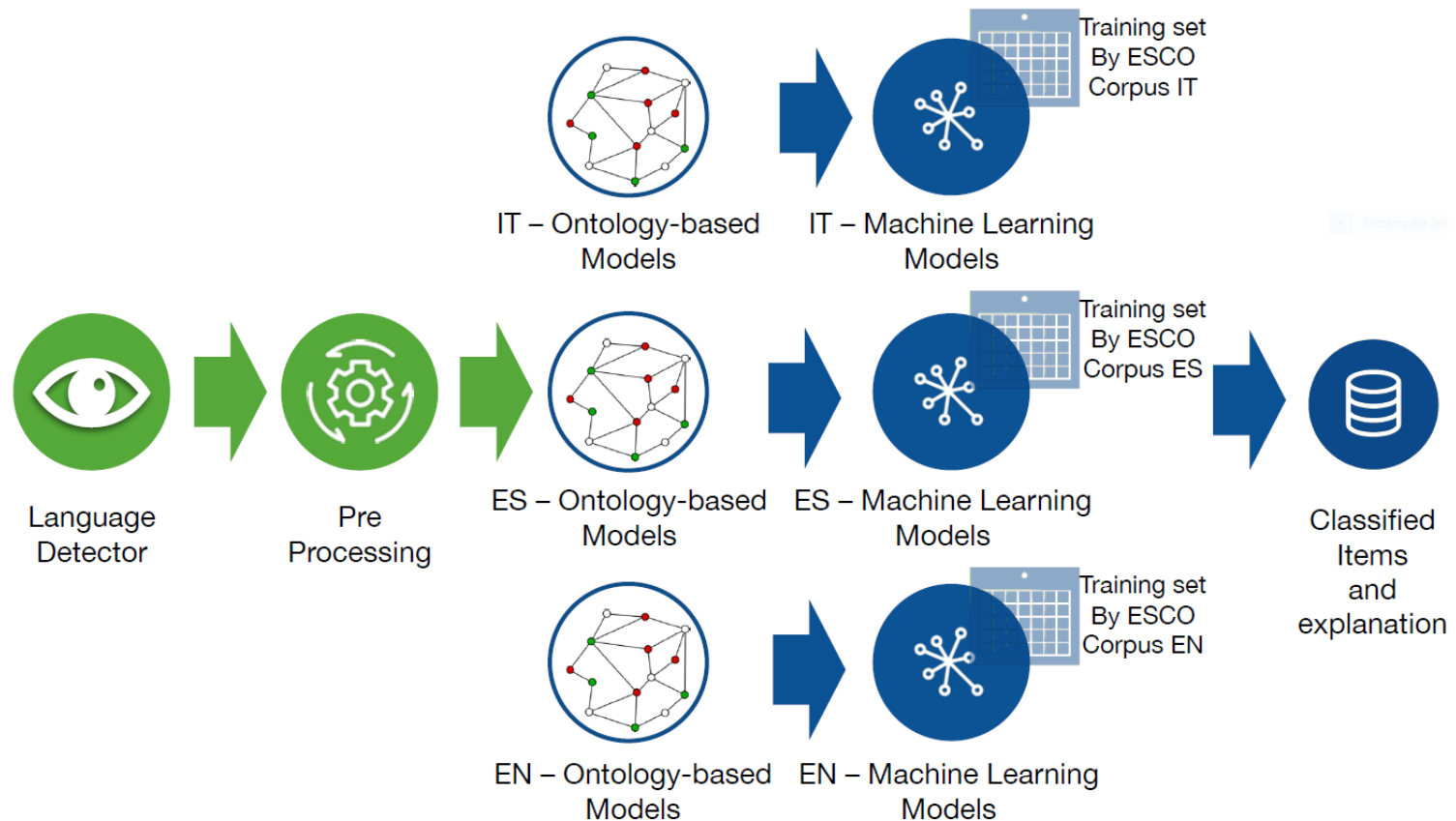
This role will directly support the Head of Data Labs in the design and implementation ...

# Data pipeline

Landscaping



# Extracting data from an OJA natural language

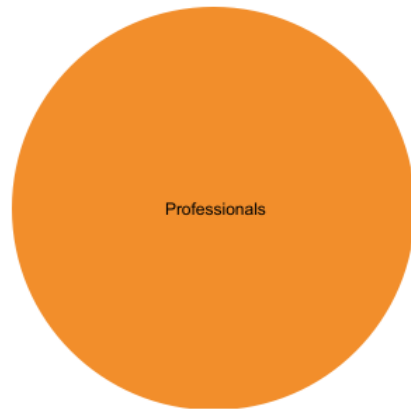


# Extracting data from an OJA (Cedefop)

| Variable          | From structured fields | From natural language | Overall precision rate (natural language) |
|-------------------|------------------------|-----------------------|---|
| Occupation        | 4.74%                  | 100.00%               | 85%                                       |
| Education level   | 4.50%                  | 99.86%                | 99%                                       |
| Experience        | 2.68%                  | 45.64%                | n.a.                                      |
| Contract          | 29.76%                 | 79.64%                | 99%                                       |
| Economic activity | 39.92%                 | 100.00%               | 92%                                       |
| Working hours     | 24.14%                 | 64.66%                | 99%                                       |
| Place             | 88.42%                 | 97.86%                | n.a.                                      |
| Salary            | 23.32%                 | 21.38%                | n.a.                                      |
| Skill             | 0.81%                  | 90.43%                | n.a.                                      |
| Release date      | 100.00%                | 100.00%               | n.a.                                      |
| Expire date       | 13.18%                 | 100.00%               | n.a.                                      |

# Main skills required for Office professionals

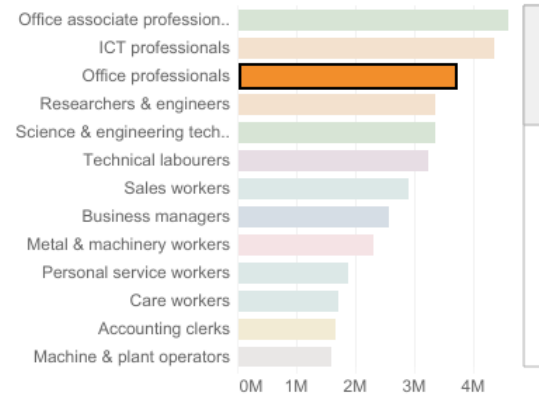
Online job vacancies by occupation (1 digit ISCO)



Most requested skills (ESCO version 1)

○ Treemap  
● Word cloud

Online job vacancies by occupation (2 digit ISCO)



Minimum share of vacancies where skill is mentioned  
22.00% to 100.00%

Online job vacancies by occupation (4-digit ISCO)



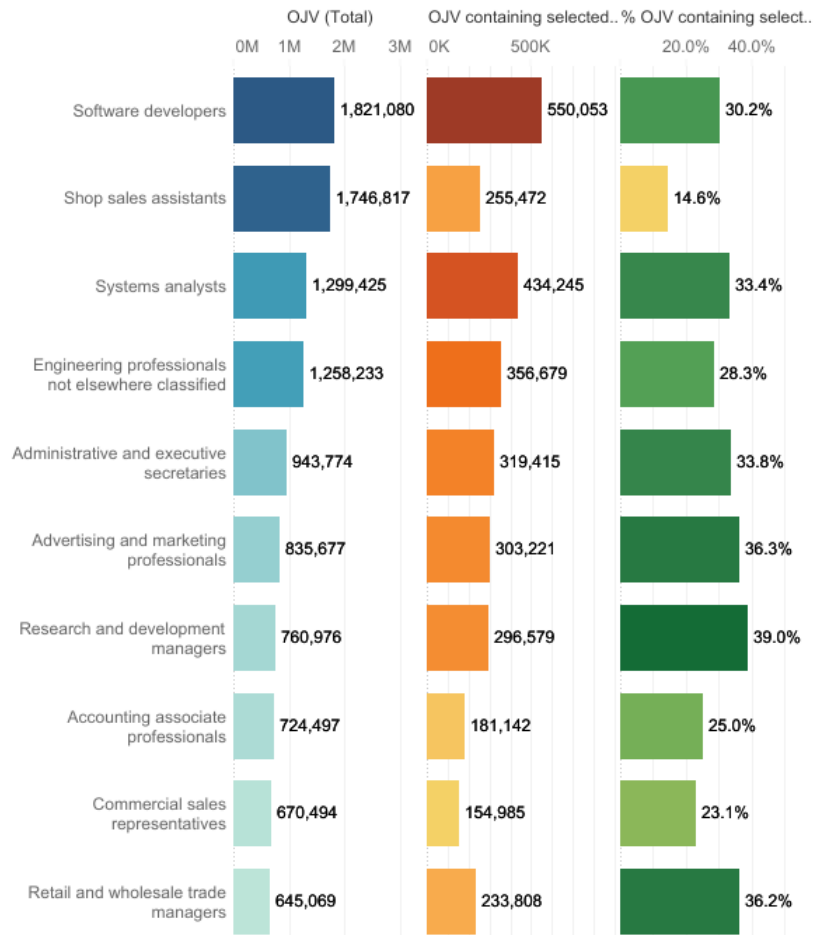
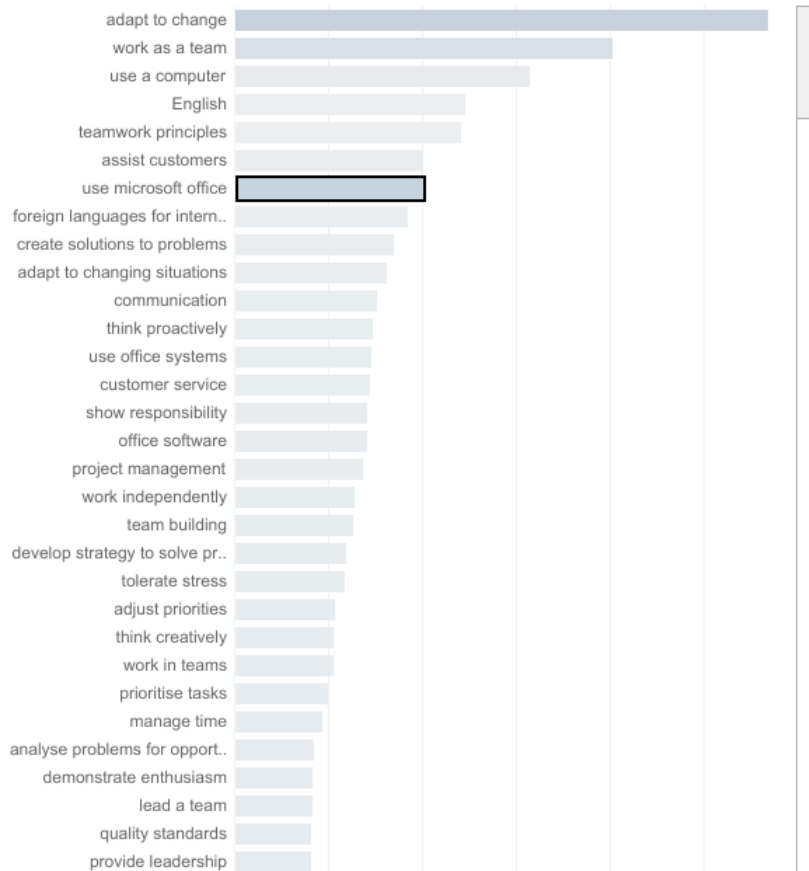




European  
Commission

# Occupations for which the skill “Use Microsoft office” is most frequently required

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



# Enhanced job vacancies statistics

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

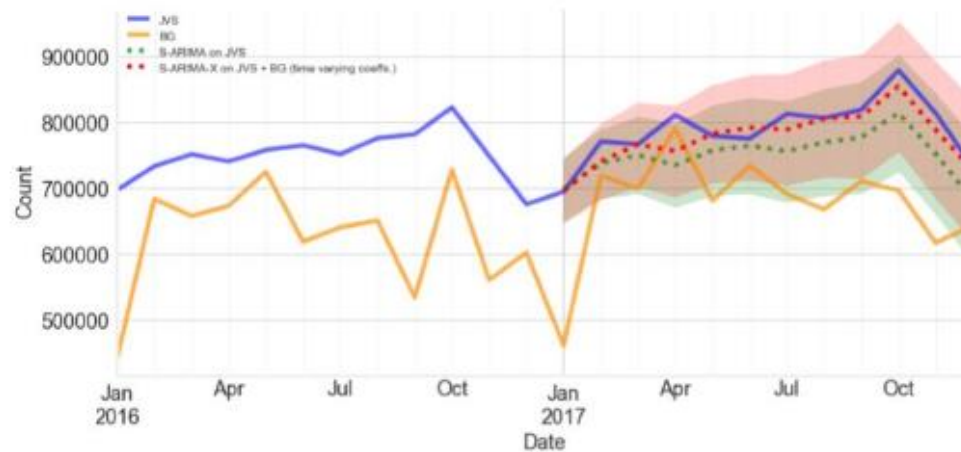
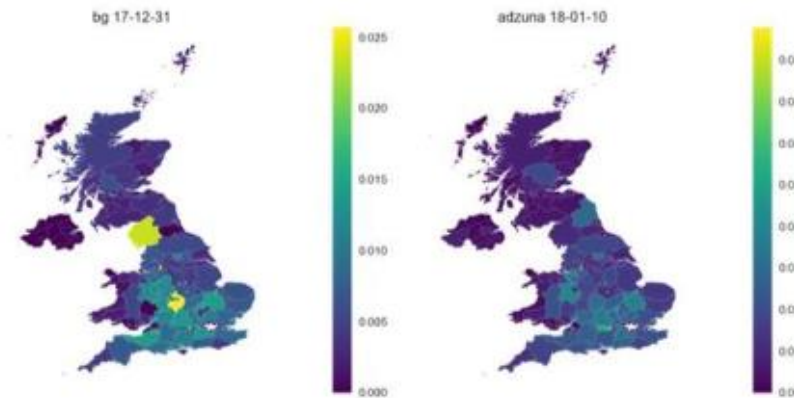


Figure 6: Number of job vacancies as a proportion of working age population<sup>28</sup>



# Web Intelligence Hub – Main milestones

- Launching phase: 2019
  - **Agreements with Cedefop for OJA: Dec 2020**
- Project phase: 2020 – 2023
  - **WIH base system 1<sup>st</sup> 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**