The Web Intelligence Hub and opportunities in the context of COVID-19

Fernando Reis
Methodology and innovation unit
Eurostat
The Web as a statistics data source

- CoViD19 confinement led to the stop of many statistical operations
- Is the Web an alternative?
- The Web is more a complement than a replacement for more traditional sources!
- But: there are some examples
  - Online prices, ICT
The Web as a statistics data source

- Web scrapping is easy, however...
- You want it to be:
  - Automated
  - Methodologically sound
  - Robust
  - Transparent
  - Reproducible
  - Consistent
  - Efficient
  - Time-series
The Web as a statistics data source

- Web scrapping is easy, however...
- Producing official statistics is difficult!

- The WIH is our tool to take care of the difficult part.
From Web data to statistics pipeline

1. **Ingestion**
   - Data Ingestion
     - URLs
     - Raw unstructured data

2. **Processing**
   - Pre-Processing
     - Natural language
   - Information Extraction
     - Classified microdata

3. **Front end**
   - ETL
     - Statistics

**Landschapend**
Web Intelligence Hub - Principles

Some principles

- ESS hub
- Serving national and European needs
- Modular structure
- Defined processes and products to be guaranteed
- Priority to working together, possibility to act individually
- Programs should be open source
- 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 (WIH)

- WIH = Data + Methodologies + Platform + People + Training + Governance + Communication
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
  - Methodological guidelines
  - R&D collaboration
  - Regulatory aspects
  - IT infrastructure and tools
  - Analytical services (e.g. NLP)
  - Skills (training material)
  - Governance
How is the WIH organised?

Statistics layer
- Skills statistics
- Augmented job vacancies statistics

Application layer
- Online Job Advertisements
- Multinational enterprises business data

Services layer
- Web Intelligence Platform

Data layer
- Web data

Use Case 1
- Skills statistics
- Augmented job vacancies statistics

Use Case 2
- Augmented EGR
- Multinational enterprises business data

New Use Case
- Enhanced prices statistics
- Online prices
How to create a WIH community

• Web Intelligence Network (WIN)
  • Promote use of WIH in official statistics production
  • Support its methodological development;
  • Operationalise the role of the NSIs in the WIH:
    ✓ use the services of the WIH;
    ✓ participate in the running of the use cases;
    ✓ contribute to the development of the services;

• Call to NSIs open until 17 November
Roadmap

1. Create IT infrastructure for the WIHP
   Done

2. Deploy OJA Cedefop system to WIHP IT infrastr.
   Starting now

3. Design architecture for reusability for WIHP
   Draft done, discussing with stakeholders

4. Develop initial components for WIHP
   Starting end of this year

5. First experimental statistics based on OJA
   End of this year

6. Porte OJA to WIHP
   2021

7. Launch new use case(s) in WIHP
   2021
Thank you for your attention

Fernando Reis

Eurostat Unit Methodology and Innovation

fernando.reis@ec.europa.eu
https://github.com/reisfe/
https://twitter.com/reisfe/
https://linkedin.com/in/reisfe/