

Web Intelligence in Practice. How to use content from the web for enterprise statistics?

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829



**Web Intelligence
Network**



**Funded by
the European Union**



Join at
slido.com
#1886 116

Is web data used in your organisation?

Yes, as experimental statistics

0%

Yes, already in the production of official statistics

0%

No

0%





Join at
slido.com
#1886 116

Have you ever heard of the Web Intelligence Hub or Web Intelligence Network?

Yes, I heard about the Web Intelligence Hub
 0%

Yes, I heard about the Web Intelligence Network
 0%

No, I haven't heard about neither one
 0%



Introduction to WIH and WIN (Web Intelligence Hub) (Web Intelligence Network)

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829



**Web Intelligence
Network**



**Funded by
the European Union**

The Web as a statistics data source

- Web scrapping is easy, however...
- You want it to be:
 - Automated
 - Robust
 - Methodologically sound
 - Transparent
 - Reproducible
 - Consistent
 - Efficient
 - Comparable over time



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.

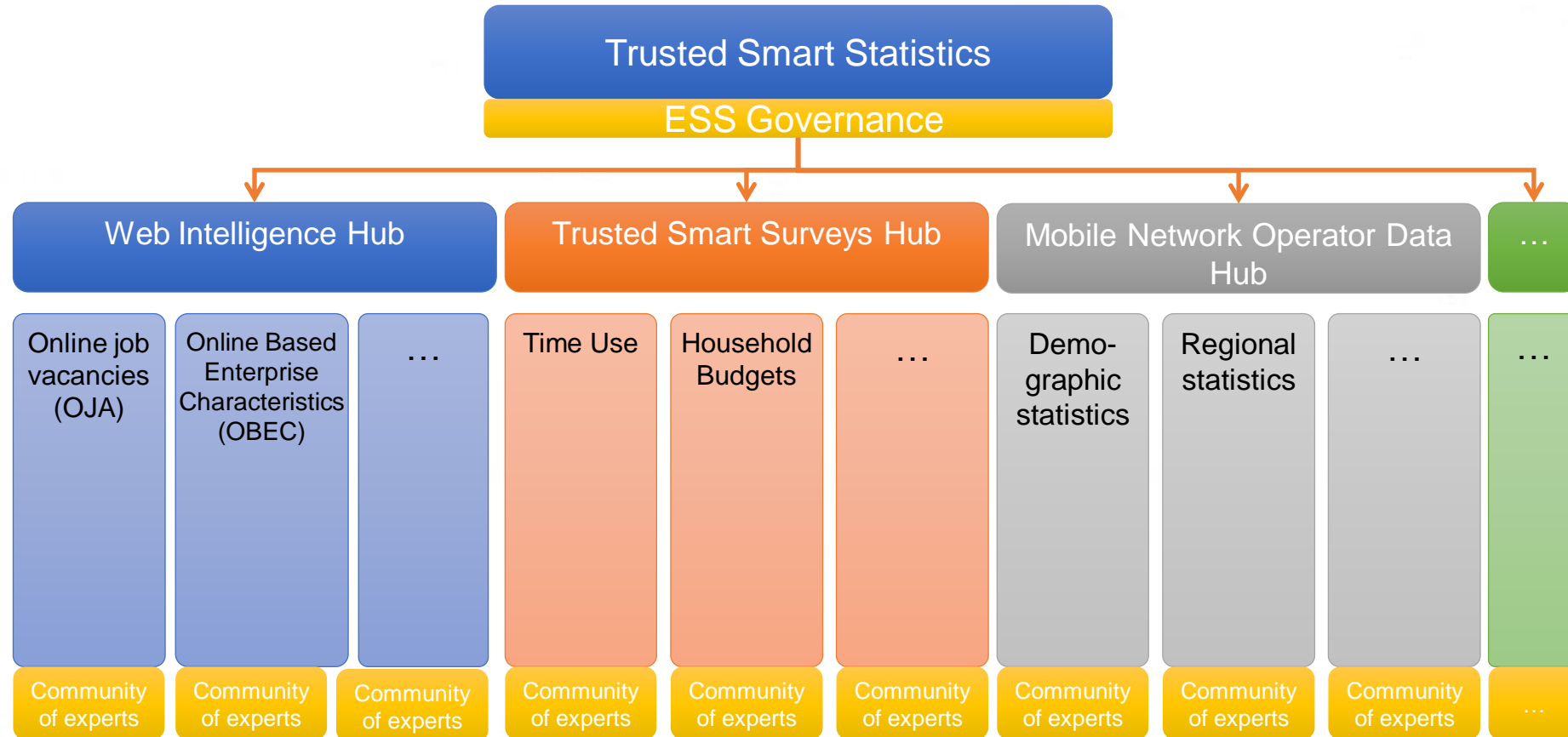


Web Intelligence
Network

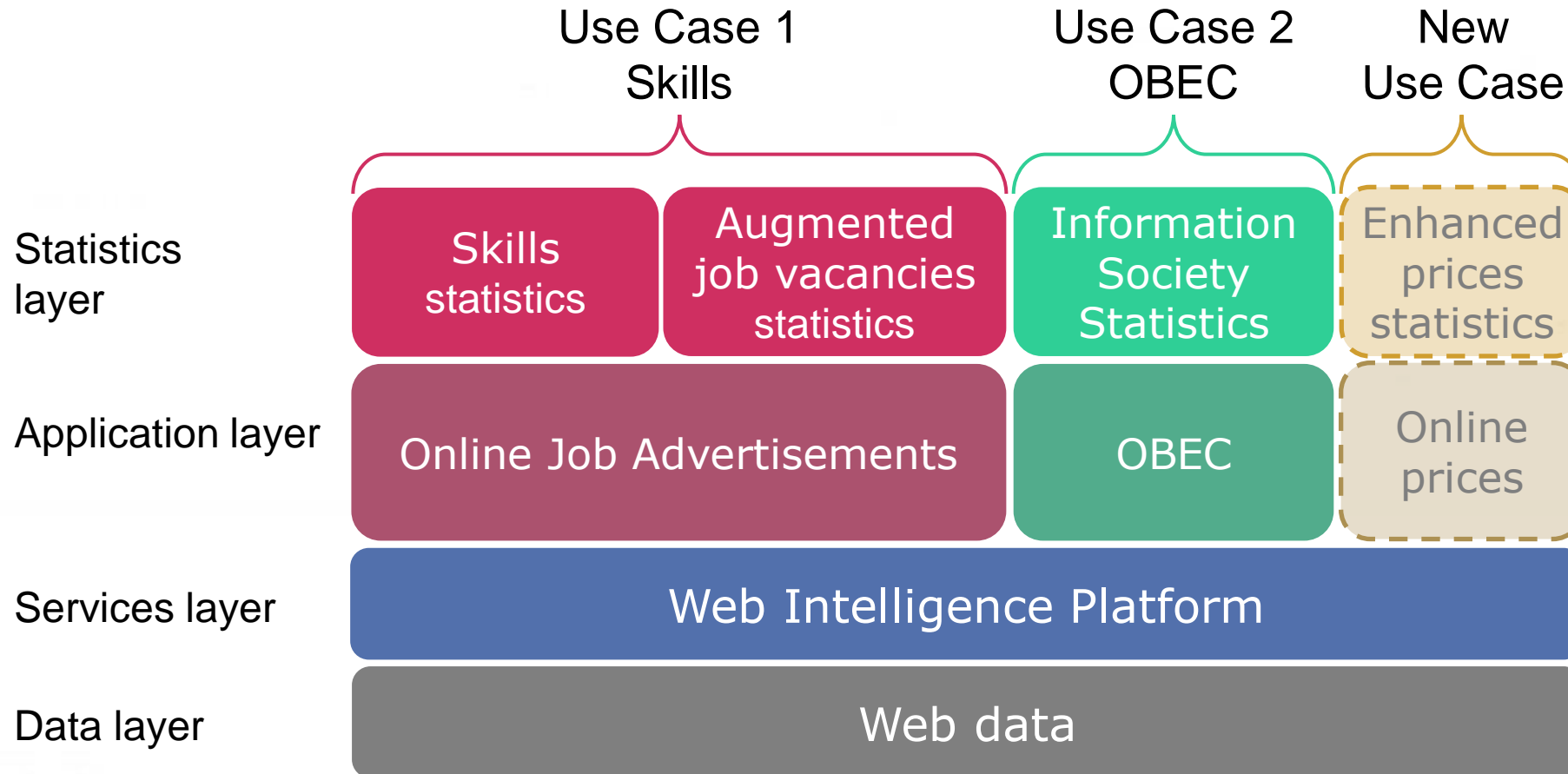


Funded by
the European Union

Trusted Smart Statistics



Web Intelligence Hub



Web Intelligence Hub Principles

ESS joint infrastructure

- Serving national and European needs
- Priority to working together, possibility to act individually
- Intermediate data products usable by all partners

Transparency

- Open source
- Commonly used processes should be certified and audible

Technically robust

- Modular architecture
- Defined processes and products guaranteed
- Clear lineage of data and processes



Web Intelligence Hub Services

- Statistical production
 - Content retrieval from websites
 - Web content processing
 - Analytical services (e.g. natural language processing)
- Business support functions
 - IT infrastructure
 - Content provision agreements and partnerships with website owners
 - Agreed methodologies & Best practices
 - Identification of regulatory and institutional needs
 - Training material for capacity building
 - R&D collaboration
 - Governance



Web Intelligence
Network



Funded by
the European Union

Building the WIN. Community of WIH users and contributors

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

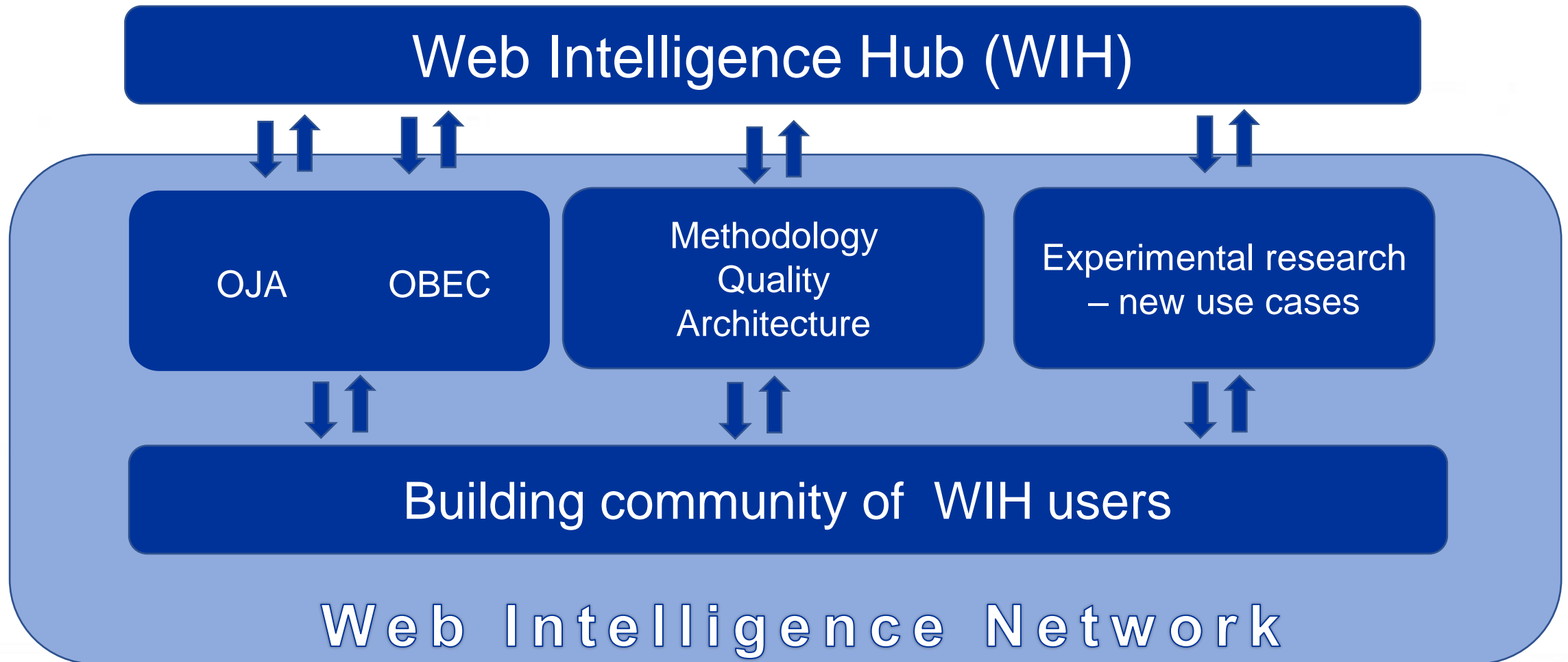


**Web Intelligence
Network**

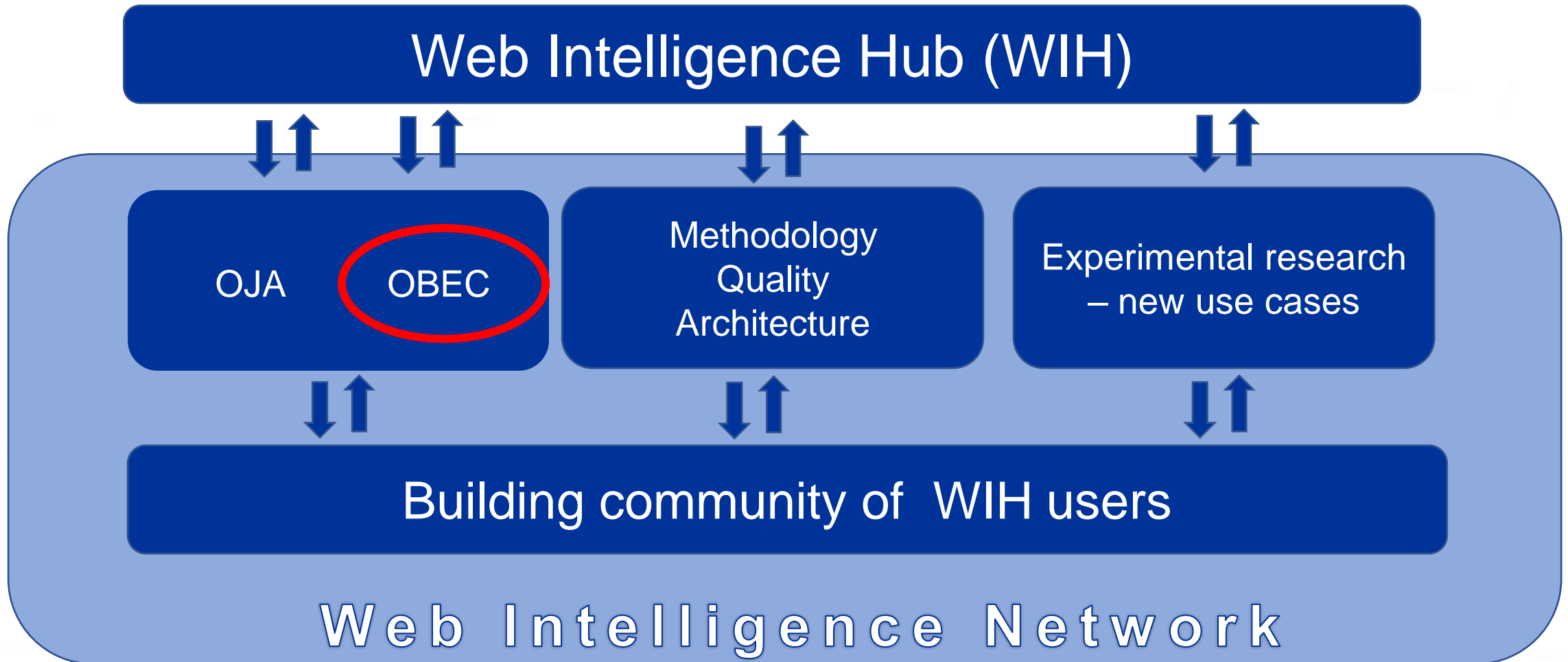


**Funded by
the European Union**

What is the Web Intelligence Network (WIN)?



What is the Web Intelligence Network (WIN)?

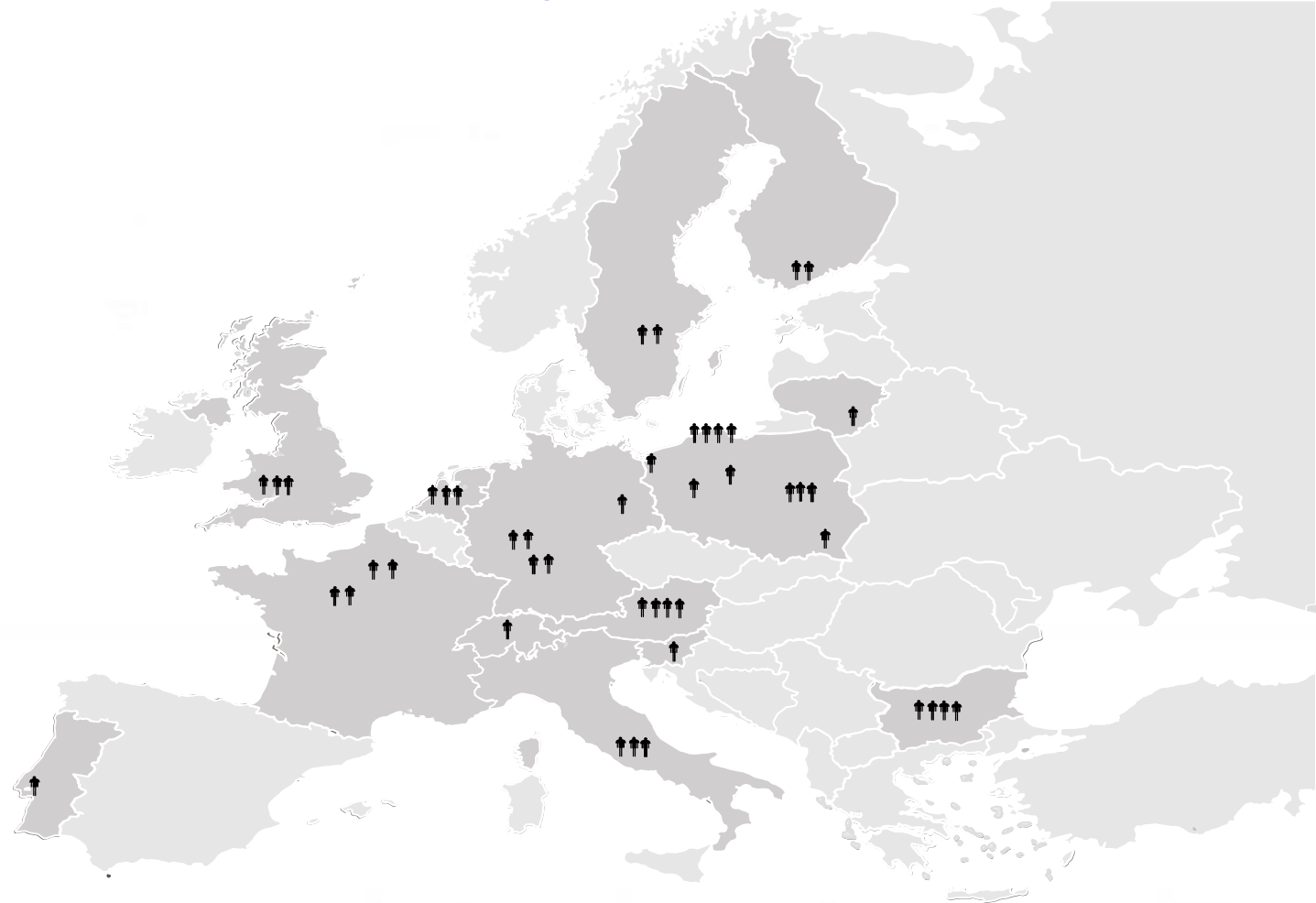


Building the WIN – a community of WIH users

Reach out to all ESS countries



Use web data, use the WIH



Web Intelligence
Network



Funded by
the European Union

Building the WIN – a community of WIH users

Follow our progress by:

- Joining the [WISER Group](#)
- Visiting our website: https://ec.europa.eu/eurostat/cros/WIN_en
- Read our [blogs](#)
- Follow us on [Twitter](#)
- Follow us on [LinkedIn](#)
- Contact us: do.nowak@stat.gov.pl



Web Intelligence
Network



Funded by
the European Union



Join at
slido.com
#1886 116

Does your organisation use web data for enterprise statistics?

Yes
 0%

No
 0%

Not sure
 0%



Introduction to OBEC

Jacek Maślankowski
j.maslankowski@stat.gov.pl
Brussels, 10th of March 2023

Trusted Smart Statistics – Web Intelligence Network
Grant Agreement: 101035829



Web Intelligence
Network



Funded by
the European Union

Agenda



What is OBEC?

How can we define OBEC population?

What is the reason to include enterprises having 10+ employees?

What is the value added to official statistics?

What experimental statistics on OBEC have already been disseminated?

How WIN supports OBEC?



Web Intelligence
Network



Funded by
the European Union



Goal of the presentation

Overview of what can be measured with the use of Big Data methods to support official statistics in terms of Online Based Enterprise Characteristics (OBEC).

Sources used in this presentation:

- ESSnet Big Data I deliverables
(https://ec.europa.eu/eurostat/cros/content/wp2-documentation1_en)
- ESSnet Big Data II deliverables
(https://ec.europa.eu/eurostat/cros/content/WPC_Enterprise_characteristics_en)
- Web Intelligence Network deliverables
(accessed via WIN wiki and blog)

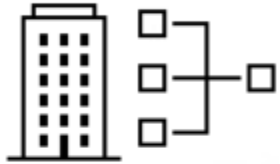


Web Intelligence
Network



Funded by
the European Union

What is OBEC?



Online Based Enterprise Characteristics



The use of a website by the enterprise to present its 'business', with extension to social media perspective.



It includes not only the existence of a website which is located on servers belonging to the enterprise or at one of the enterprise's sites, but also third party's websites (e.g. one of the group of enterprises to which it belongs, i.e. the website of the parent enterprise).



How can we define OBEC population?



Methodological manual for data compilers and users of the ICT survey:

A4. Does your enterprise have a website?

[Scope: enterprises with access to the internet, i.e. A1 > 0],

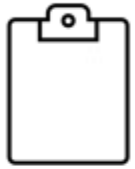
[Type: single answer (i.e. Tick only one); binary (Yes/No); filter question].



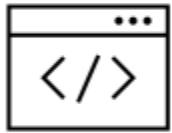
The population of OBEC use case consists of enterprises having website and employing 10 or more employees.



What is the reason to include enterprises having 10+ employees?



Traditional questionnaire – Survey on ICT Usage and E-commerce in Enterprises



URL database (Uniform Resource Locator – website address)



Significant percentage of small enterprises does not have a website compared to larger enterprises



Names of small companies are usually not unique



Question?

- Please consider asking enterprises in selected countries about their e-commerce activity.
- Which information is more reliable?
 - a) based on traditional representative survey
 - b) based on the analysis of the content of all websites identified as enterprise websites in the country



What is URL database?

It should consist of two datasets:

- For Web Intelligence Platform (web scraping tool with Elasticsearch):
 - Anonymized ID
 - Enterprise URL
 - Group to which it belongs (e.g. /OBEC)
- For further processing in Datalab (JupyterLab with Python, RStudio):
 - Business register ID
 - Anonymized ID
 - Other attributes

To link with business register



Web Intelligence
Network



Funded by
the European Union

Question?

- Do you maintain any URL database in your institution, e.g. business register additional column or experimental database/file?
 - a) yes
 - b) no



URL database. Where to find URLs of enterprises?

Country	Database name
AT	Unofficial URL list obtained during ESSNet BD II
BG	1. BNSI database, a non-official list with obtained URLs under the ESSnet on BD I and II projects (approx 13k) 2. BNSI SBR data (approx 2k)
DE (Destatis)	URLs are currently being acquired from a third-party vendor. Delivery of URLs expected.
DE (HSL)	2k manually searched enterprise URLs (retail sector, done in 03/2021)
IT	1. Unofficial URL list obtained in the past years within a research project (approx 80k)
LT	1. Statistical Economic Entities register (approx. 5K URLs) 2. Rekvizitai.lt database (approx. 9K URLs)
PL	1. REGON business register, Statistics Poland (>200K. URLs) 2. BJS business register, Statistics Poland (>200KURLs) 3. CEiDG, government institution (unknown) 4. KRS, government institution (unknown) 5. ORBIS, external company (>500K URLs)
CH	1) Swiss Business Register 2) Other potential sources still under scrutinization



What is the value added to official statistics?

Case study on ICT Usage and E-commerce in Enterprises

COMMUNITY SURVEY ON ICT USAGE AND E-COMMERCE IN ENTERPRISES

2023

General outline of the survey

Web presence		
Use of a website		
A4. Does your enterprise have a website? (Filter question)	Yes <input type="checkbox"/>	No <input type="checkbox"/> <small>-> go to A6</small>
A5. Does the website have any of the following?	Yes	No
a) Description of goods or services, price information	<input type="checkbox"/>	<input type="checkbox"/>
b) Online ordering or reservation or booking, e.g. shopping cart	<input type="checkbox"/>	<input type="checkbox"/>
c) Possibility for visitors to customise or design online goods or services	<input type="checkbox"/>	<input type="checkbox"/>
d) Tracking or status of orders placed	<input type="checkbox"/>	<input type="checkbox"/>
e) Personalised content on the website for regular/recurrent visitors	<input type="checkbox"/>	<input type="checkbox"/>
f) A chat service for customer support (a chatbot, virtual agent or a person replying to customers)	<input type="checkbox"/>	<input type="checkbox"/>
g) Advertisement of open job positions or online job application	<input type="checkbox"/>	<input type="checkbox"/>
h) Content available in at least two languages <small>Please, consider a multilingual website within a single domain (e.g. ".com") or multiple domains of your enterprise in different languages (e.g. ".es", ".uk").</small>	<input type="checkbox"/>	<input type="checkbox"/>

Use of social media		
Enterprises <u>using</u> social media are considered those that have a user profile, an account or a user licence depending on the requirements and the type of the social media.		
A7. Does your enterprise use any of the following social media? <i>(add national examples; replace existing examples if necessary)</i>	Yes	No
a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer)	<input type="checkbox"/>	<input type="checkbox"/>
b) Enterprise's blog or microblogs (e.g. Twitter)	<input type="checkbox"/>	<input type="checkbox"/>
c) Multimedia content sharing websites or apps (e.g. YouTube, Flickr, SlideShare, Instagram, Pinterest, Snapchat)	<input type="checkbox"/>	<input type="checkbox"/>

A8. Does your enterprise use any of the above mentioned social media to: - optional		
	Yes	No
a) Develop the enterprise's image or market products (e.g. advertising or launching products)	<input type="checkbox"/>	<input type="checkbox"/>
b) Obtain or respond to <u>customer</u> opinions, reviews, questions	<input type="checkbox"/>	<input type="checkbox"/>
c) Involve <u>customers</u> in development or innovation of goods or services	<input type="checkbox"/>	<input type="checkbox"/>
d) Collaborate with <u>business partners</u> (e.g. suppliers) or <u>other organisations</u> (e.g. public authorities, non-governmental organisations)	<input type="checkbox"/>	<input type="checkbox"/>
e) Recruit employees	<input type="checkbox"/>	<input type="checkbox"/>
f) Exchange views, opinions or knowledge <u>within</u> the enterprise	<input type="checkbox"/>	<input type="checkbox"/>

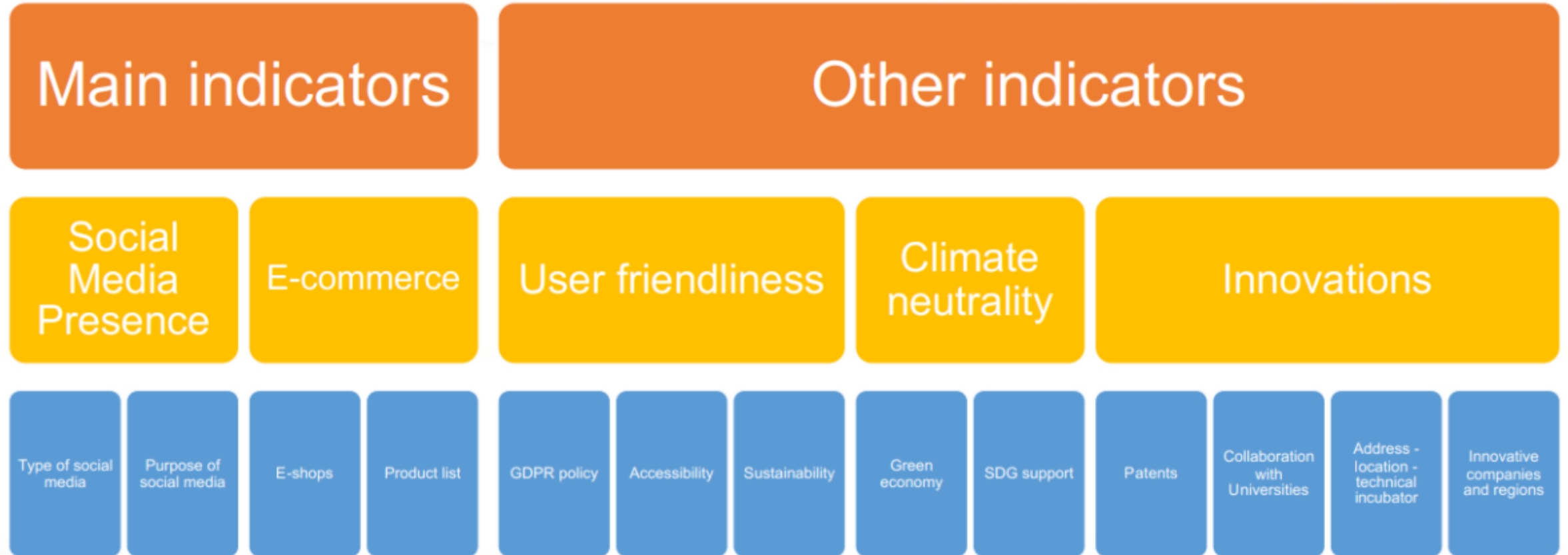


Web Intelligence
Network



Funded by
the European Union

Other website attributes – results from WIN brainstorm session



What experimental statistics on OBEC have already been disseminated?

Data was disseminated for 2018, 2019, 2020 during ESSnet Big Data I and II grants

Statistical indicators: retrieved URLs and variables in the ICT usage in enterprise survey

Statistical indicators	ICTsurvey 2020 estimates (%)	Estimates from web (%)
Rate of enterprises having websites	90.4	89.1
Rate of enterprises that are present on social media (with website)	52.1	62.1
Rate of enterprises engaged in websales on their website	34.4	33.1

Austria

Statistical indicators	Estimates from web	ICTsurvey 2020 estimates
Rate of enterprises having websites	43.5%	52.0%
Rate of enterprises engaged in websales (all enterprises)	4.7%	7.6%
Rate of enterprises engaged in websales on their website (all enterprises having a website)	10.8%	14.6%
Rate of enterprises that are present on social media (all enterprises)	8.7%	18.7%
Rate of enterprises that are present on social media (with website)	19.9%	35.9%

Bulgaria

Statistical indicators	Estimates from web	ICTsurvey 2020 Estimates
Rate of enterprises having websites	92%	92%
Rate of enterprises engaged in websales (all enterprises)		
Rate of enterprises engaged in websales on their website	14%	34%
Rate of enterprises that are present on social media (all enterprises)		
Rate of enterprises that are present on social media (with website)	76%	59%

Netherlands

Indicator	Estimates from web 2019	Estimates from web 2020	ICTsurvey 2019 estimates
Rate of enterprises having websites	73,6%	83,3%	96,3%
Rate of enterprises engaged in websales on their website (c10b)	5,1%	5,9%	13,1%
Rate of enterprises that are present in social media (c12a,b,c)	32,4%	35,3%	36,6%

Poland

Rate of enterprises that are present on social media

	Estimate from web data	Survey Estimate
Bulgaria (BNSI software)	31%	34%
Bulgaria (Polish software)	37%	34%
Italy	37%	31%
Netherlands	65%	69%
Poland	26% (Pomeranian voivodship only)	25% (Pomeranian voivodship only)
UK	80%	66%

2018

Find more here:

ESSnet Big Data I: https://ec.europa.eu/eurostat/cros/content/WP2_Experimental_statistics1_en (2018)

ESSnet Big Data II: https://ec.europa.eu/eurostat/cros/content/wpc-experimental-statistics_en (2019, 2020)



Web Intelligence Network



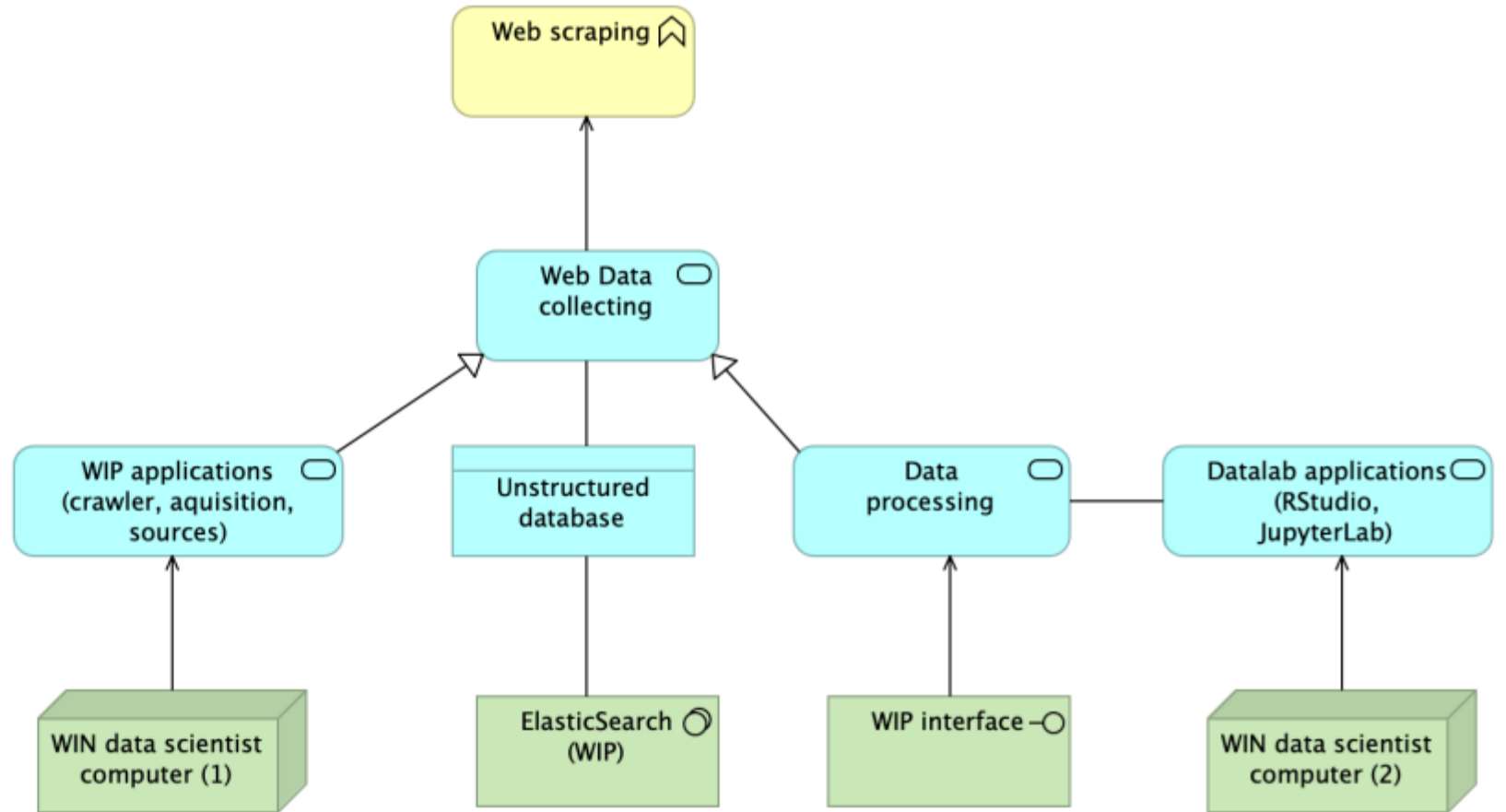
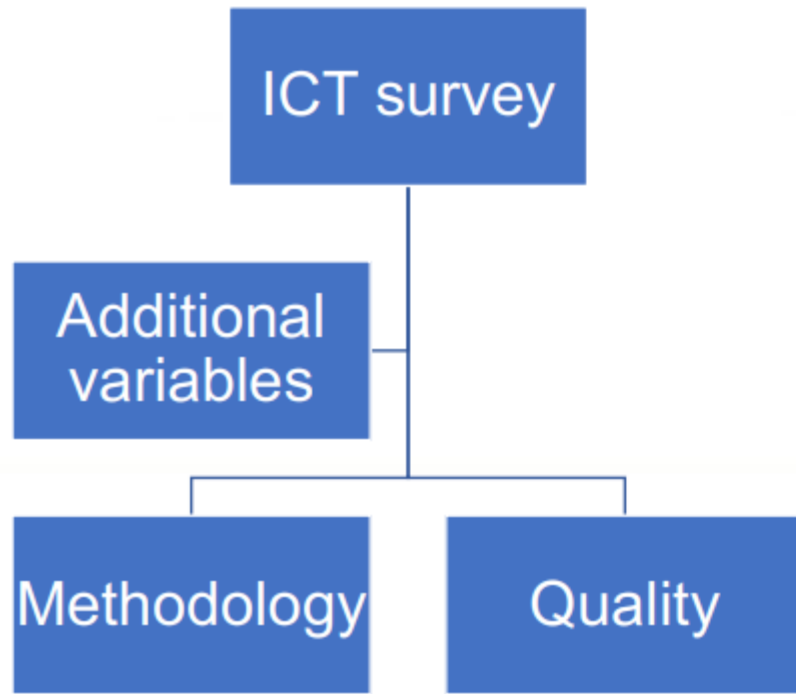
Funded by the European Union

Question?

- Consider asking about providing e-commerce activity, e.g. e-shop.
- Which data for one enterprise is more reliable – do you trust in the data based on:
 - a) the traditional questionnaire
 - b) the analysis of the content on the website



How WIN supports OBEC?



How WIN supports OBEC?

Local legislation
on enterprise
identification on
the web

Local practices
for Ecommerce
and / or social
media

Language
specifics

Re-use software

Internet stability

Internet
restrictions

Quality of
webscraping
algorithm

Allowable
response time
for a websites

URLs for the
websites

Big Data
environment

Programming
skills

Webscraping
Policy



Web Intelligence
Network



**Funded by
the European Union**

Thank you!

Jacek Maślankowski
j.maslankowski@stat.gov.pl



Web Intelligence
Network



Funded by
the European Union

Links needed for the following presentation

- <https://prod.wihp.ecdp.tech.ec.europa.eu/>
- <https://dss-dsl2531b.ecdp.dataplatform.tech.ec.europa.eu/>
- <https://git.fpfis.tech.ec.europa.eu/estat/wihp/>
- <https://estat.pages.fpfis.eu/wihp/doc-server/>





WIH platform

Overview and demo

NTTS 2023 – WIN workshop

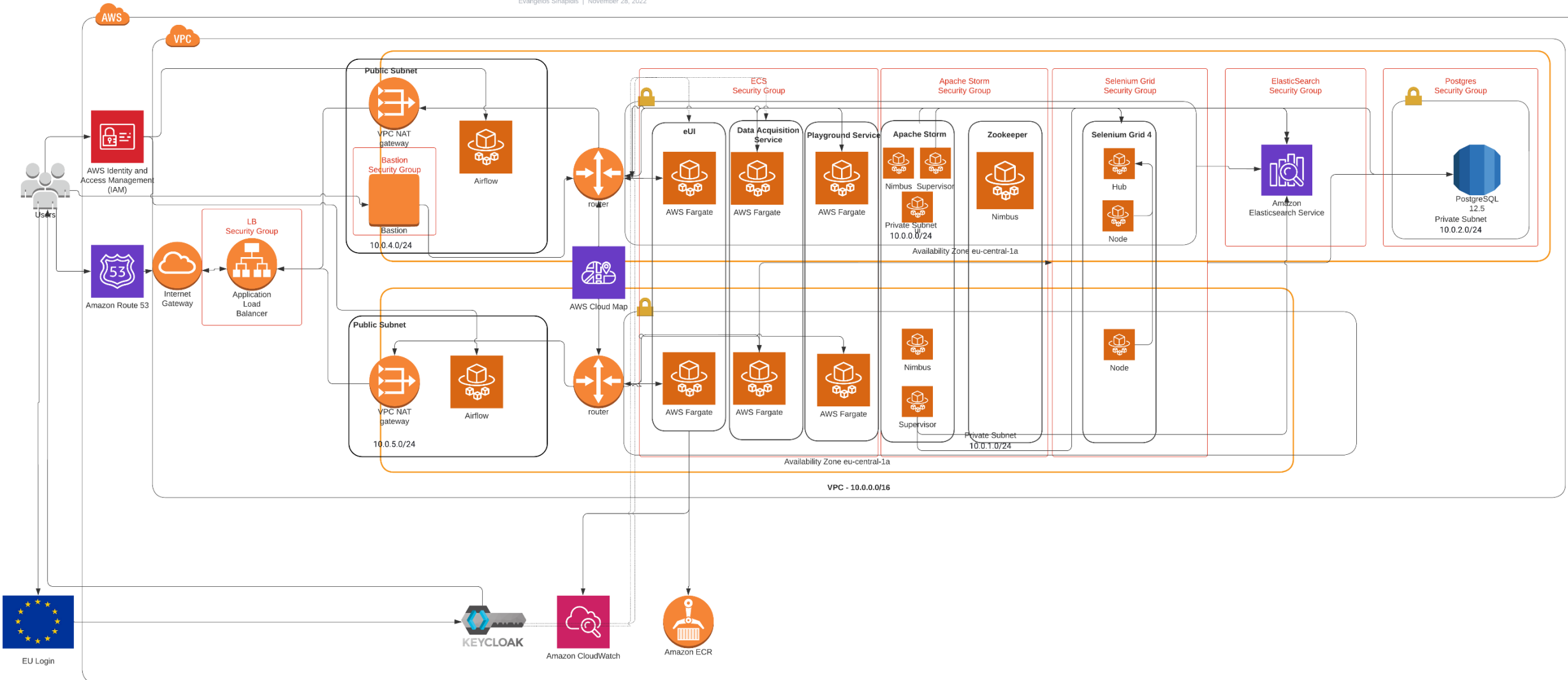
Main components

- Data Acquisition platform: <https://prod.wihp.ecdp.tech.ec.europa.eu/>
- Documentation: <https://estat.pages.fpfis.eu/wihp/doc-server/>
- GitLab CI/CD and repository: <https://git.fpfis.tech.ec.europa.eu/estat/wihp/>
- Datalab from the ECDP: <https://dss-dsl2531b.ecdp.dataplatform.tech.ec.europa.eu/>
- EU Login, Azure AD and Keycloak as ID provider, authentication and authorisation

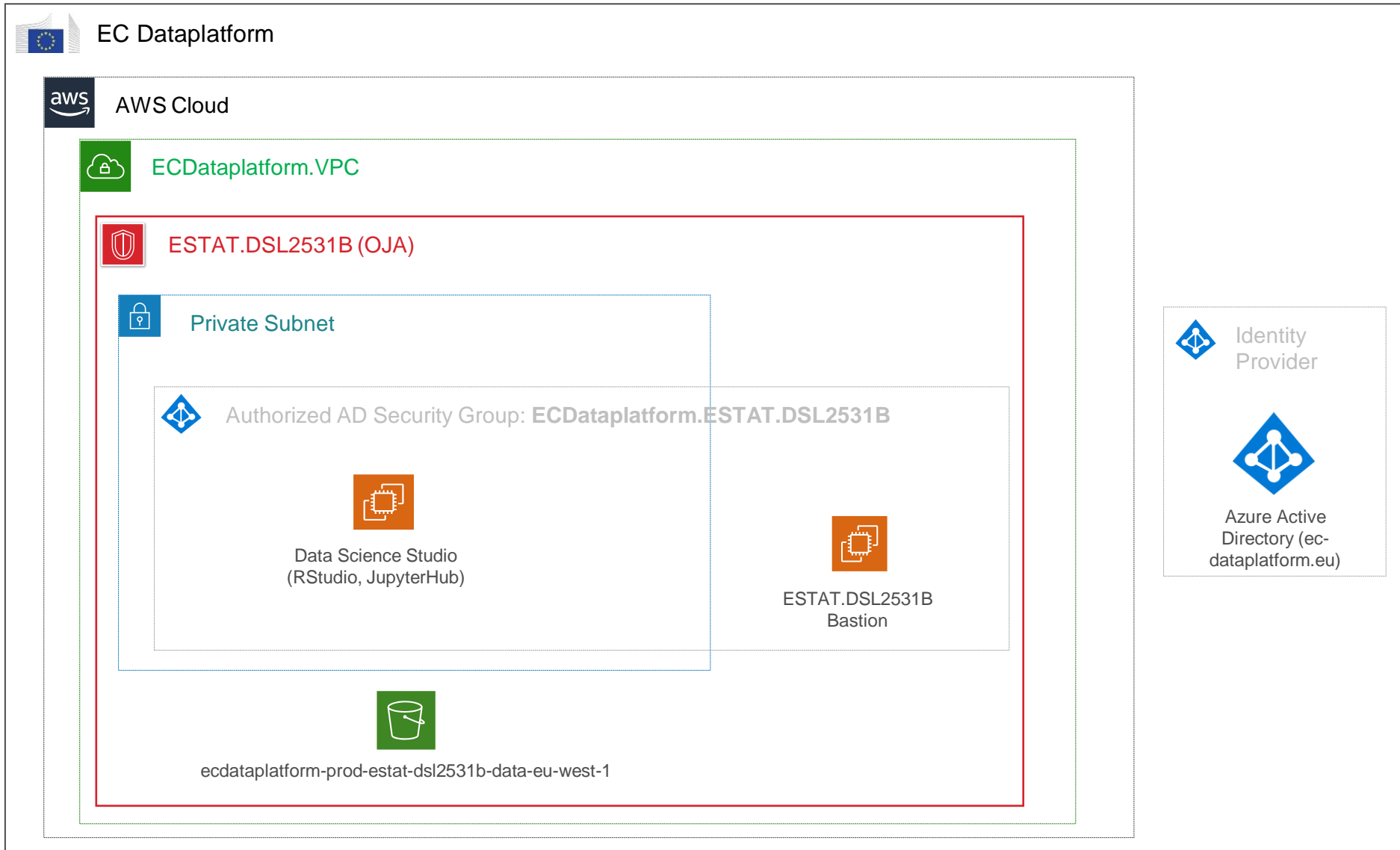
Data Acquisition

WIHP - Data Collection IaC - v3.0

Evangelos Sinapidis | November 28, 2022



Datalab



Lets see in practice...

Demo some of the features of the data acquisition platform

Thank you



© European Union 2023

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.



URL Finding

Heidi Kühnemann
Statistics Hesse

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829



**Web Intelligence
Network**



**Funded by
the European Union**

Enterprise URLs: Why?

In general:

- Freely available enterprise information on various topics
- Potential to reduce response burden
- Potential to update statistical business register with additional data source

For ICT survey:

- Enterprise websites already contain parts of the information we are interested in
 - Websites can be scraped for (nearly) the full population!
- Improve quality, reduce the number of questions and/or produce results more frequently!



Web Intelligence
Network



Funded by
the European Union

ICT-ENT survey: website

Use of a website		
A4. Does your enterprise have a website? (Filter question)	Yes <input type="checkbox"/>	No <input type="checkbox"/> ->go to A6
A5. Does the website have any of the following?	Yes	No
a) Description of goods or services, price information	<input type="checkbox"/>	<input type="checkbox"/>
b) Online ordering or reservation or booking, e.g. shopping cart	<input type="checkbox"/>	<input type="checkbox"/>
c) Possibility for visitors to customise or design online goods or services	<input type="checkbox"/>	<input type="checkbox"/>
d) Tracking or status of orders placed	<input type="checkbox"/>	<input type="checkbox"/>
e) Personalised content on the website for regular/recurrent visitors	<input type="checkbox"/>	<input type="checkbox"/>
f) A chat service for customer support (a chatbot, virtual agent or a person replying to customers)	<input type="checkbox"/>	<input type="checkbox"/>
g) Advertisement of open job positions or online job application	<input type="checkbox"/>	<input type="checkbox"/>
h) Content available in at least two languages Please, consider a multilingual website within a single domain (e.g. ".com") or multiple domains of your enterprise in different languages (e.g. ".es", ".uk").	<input type="checkbox"/>	<input type="checkbox"/>



ICT-ENT survey: social media

Use of social media		
Enterprises using social media are considered those that have a user profile, an account or a user licence depending on the requirements and the type of the social media.		
A7. Does your enterprise use any of the following social media? <i>(add national examples; replace existing examples if necessary)</i>	Yes	No
a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer)	<input type="checkbox"/>	<input type="checkbox"/>
b) Enterprise's blog or microblogs (e.g. Twitter)	<input type="checkbox"/>	<input type="checkbox"/>
c) Multimedia content sharing websites or apps (e.g. YouTube, Flickr, SlideShare, Instagram, Pinterest, Snapchat)	<input type="checkbox"/>	<input type="checkbox"/>



ICT-ENT survey: e-commerce

B1. During 2022, did your enterprise have web sales of goods or services via:	Yes	No
a) your enterprise's websites or apps? (including extranets)	<input type="checkbox"/>	<input type="checkbox"/>
b) e-commerce marketplace websites or apps used by several enterprises for trading goods or services? (e.g. e-Bookers, Booking, hotels.com, eBay, Amazon, Amazon Business, Alibaba, Rakuten, TimoCom etc.) <i>[Please add national examples of e-commerce marketplaces incl. government marketplaces]</i>	<input type="checkbox"/>	<input type="checkbox"/>



Enterprise URLs: How?

This is what we focus on today

	Obtain data from registers/survey	Data purchases	Automated procedure to search for URLs
Advantages	Low costs or freely available	Make use of extensive experience of data owner → potentially very good data quality	Costumizability Exact fit to data needs Transparent methodology
Dis-advantages	Possibly out-of-date (administrative data) Increased response burden (survey) Addition of new question to survey often not possible Incomplete	High costs Lacking methodological transparency Linkage of URLs to business register might be challenging	Time-consuming to set up Requires specialized IT-infrastructure Requires methodological, IT and programming skills



What are your experiences with enterprise websites?

Do you or your office (plan to) use enterprise websites as data source?
If yes: How do you or your office (plan to) acquire enterprise URLs?

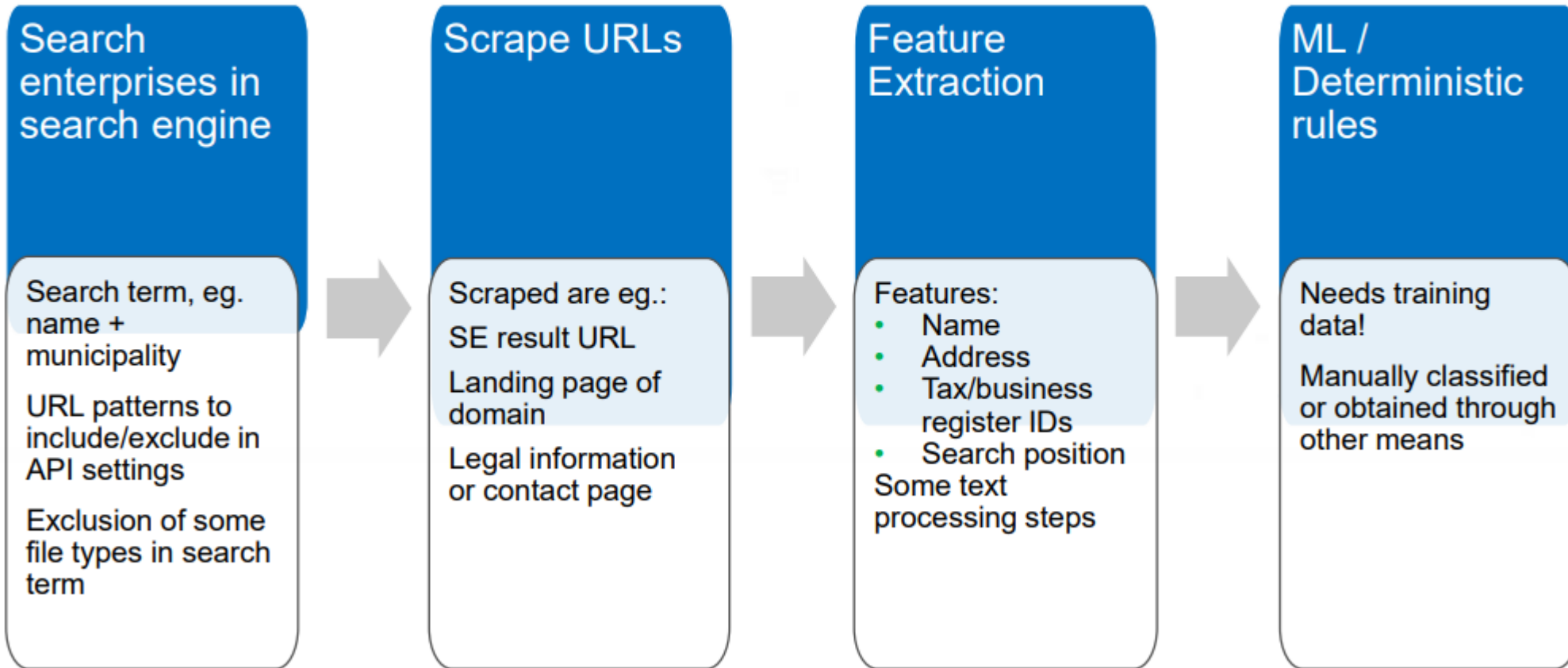


Web Intelligence
Network



Funded by
the European Union

URL finding overview



Search engines

Criteria to consider when selecting a search engine:

- Can a SE identify the correct URLs?
- Limits in the number of requests
- Costs of requests

% domains matched	GOOGLE	GOOGLE API	BING	YAHOO	DUCK
Italian sample	74.8	66.7	64.7	63.6	57.6
Hessian sample	89	87	62	59	NA

Comparison of SE results for ca. 100 Italian and Hessian enterprises



API or Search Engine Scraping?

API:

- ✓ Many configuration options
- ✓ High frequency of requests possible
- ✗ Only small number of requests are free

Search Engine Scraping:

- ✓ Requests are free
- ✓ Obtain results like a human being
- ✗ Potential violation of terms of use
- ✗ Scrapers might get blocked



What are your experiences with search engines?

Have you used a search engine as data source before (either by scraping results or via API)?

For what kind of purposes have you used search engine results?



Web Intelligence
Network



Funded by
the European Union

Scraping

- By far the most cumbersome step: scrape all result URLs
- Each search produces ca. 10-30 URLs to be scraped (result URLs, contact pages, imprint, landing,...)
- URLs are very diverse: different technologies, sometimes large contents
- Information is sometimes hidden in Javascript → Javascript rendering software is advisable (automated browser)
- Headless browsers: Selenium or Splash are in use within ESS
- But: Javascript rendering increases the amount of downloaded data and bandwidth usage
- Massive scraping needs special infrastructure



Feature Extraction

- Preprocessing steps, eg.
 - remove css styles and javascript code
 - remove duplicate whitespaces
 - lowercasing words and letters
- Compare enterprise data from SBR with scraped data, eg.
 - Name is on website
 - VAT ID is on websites,
 - ...
- Features are created with exact string matching or regular expressions
- String similarity for comparison of short texts with enterprise data (eg. name and HTML title)



Machine Learning / Deterministic Rules

- When do we accept a URL as correct?
- Deterministic rules:
 - eg. VAT ID on website → website correct
 - Easy to build and interpret
 - What if enterprise data is missing in the SBR or on the website?
 - What if other website mentions data of different enterprises?
 - Validation data necessary to measure classification performance
- Machine Learning:
 - Training & validation data necessary
 - Model decides which features have which weight
 - Reduced interpretability



Differences in approaches

- **No scraping:** Search engine results are used directly, as done by:
 - NL: Search engine data & snippets are used for feature engineering
 - FI: Extensive URL exclusion/inclusion rules are applied
- **External data sources:** domain registration data (FI)
 - Dataset with all .fi domains including information on owner (eg. business name & ID)
 - Use data eg. instead of search engine results



Challenges

- Training data is often not available in sufficient quality
 - Evaluation and machine learning approaches not possible/reliable or ...
 - time-consuming manual search of URLs necessary
- Scalability and resource demands
 - Downloading, storing and analyzing scraped URLs takes time and processing power
→ Only use search engine results (no scraping) or ...
 - make parallel processing possible
- Legal questions, eg. statistical confidentiality when using search engines



What is your opinion/experience?

What challenge do you consider the most crucial to solve when planning to implement URL finding in your own organisation?

It can be challenges already mentioned or others.



Web Intelligence
Network



Funded by
the European Union

URL finding on the WIHP?

- URL finding requires enterprise data to obtain search results and to identify correct URLs
- Enterprise data cannot be uploaded to the WIHP at the moment (data security, organisational obstacles, etc.)
- Scraping search engines or receiving search engine API results is currently not part of the WIHP functionality
- URL finding is not implemented on the WIHP at the moment
- But: The search engine part can be done locally, and the result URLs could be scraped in the WIHP



URL finder software

- Statistics Netherlands (Python): <https://github.com/SNStatComp/urlfinding>
- Statistics Bulgaria (Python):
<https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/tree/master/URLsFinder>
- Istat (Java):
<https://github.com/EnterpriseCharacteristicsESSnetBigData/UrlSearcher>
- Additional URL finders were developed by Statistics Hesse (R), Statistics Austria (R) and Statistics Finland (Python) but have not been published yet



How well does it work? – Evaluation scores

URL finder	F1 score (level: websites)	Evaluation score „of choice“
BNSI	0.59	Precision: 0.72
Istat	0.81	Accuracy: 0.79; F1: 0.81
Statistics Hesse	0.84 (2021: 0.82)	88.2% of enterprises correct (2021: 82.3%)
CBS	0.77	F1 score on enterprise level: 0.84



Conclusions

- URL finding performs sufficiently well in most cases
- Methodological improvements are ongoing
- URL finding has been developed in different institutes with different approaches due to eg.:
 - country- and language specific particularities, eg. imprint in German speaking countries
 - experience with different programming languages in statistical offices
 - different IT infrastructures and scraping/crawling software already in use
 - existing URL finders were too slow or too computationally expensive to be run on a large scale



Literature / Further reading

- WIN Report on URL finding methodology (https://ec.europa.eu/eurostat/cros/system/files/20220131_url_finding_methodology.pdf)
- Delden, Arnout van; Windmeijer, Dick; Bosch, Olav ten (2019): Searching for business websites. CBS (Discussion Paper). <https://www.cbs.nl/engb/background/2020/01/searching-for-business-websites>.
- Barcaroli, Giulio; Scannapieco, Monica; Summa, Donato (2016): On the Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web. In: Italian Review of Economics, Demography and Statistics 4 (70), S. 25–41. http://www.sieds.it/listing/RePEc/journal/2016LXX_N4_RIEDS_25-41_Scannapieco.pdf



**OBEC step by step. Live demonstration. Jacek Maślankowski
j.maslankowski@stat.gov.pl Brussels,
10th of March 2023**

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829



**Web Intelligence
Network**



**Funded by
the European Union**

Prerequisites.

What we need to create statistics with WIN?

- Access
 - WIN WIP (<https://prod.wiwp.ecdp.tech.ec.europa.eu/screen/home>)
 - WIN Datalab (<https://dss-dsl2531b.ecdp.dataplatform.tech.ec.europa.eu>)
- Scripts
 - https://github.com/jmaslankowski/WP2_OBEC_Starter
 - JSON file with URLs for web scraping (optional)
- Skills (basics)
 - Python
 - ElasticSearch
 - Linux



Web Intelligence
Network



Funded by
the European Union

Collecting and processing data in 6 steps

1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

4. Opening WIP – cloning the Github open code on SMP.

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.



Collecting and processing data in 6 steps

1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

4. Opening WIP – cloning the Github open code on SMP.

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.



1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

```
{
  "sources": [
    {
      "name": "PL_00000001",
      "url": "https://www.stat.gov.pl",
      "group": "/OBEC"
    },
    {
      "name": "PL_00000002",
      "url": "https://ug.edu.pl",
      "group": "/OBEC"
    },
    {
      "name": "PL_00000003",
      "url": "https://gdansk.stat.gov.pl",
      "group": "/OBEC"
    },
    {
      "name": "PL_00000004",
      "url": "https://szczecin.stat.gov.pl",
      "group": "/OBEC"
    },
    {
      "name": "PL_00000005",
      "url": "https://wzr.ug.edu.pl",
      "group": "/OBEC"
    }
  ]
}
```

```
{
  "name": "PL_00000001",
  "url": "https://www.stat.gov.pl",
  "group": "/OBEC"
}
```



Collecting and processing data in 6 steps

1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

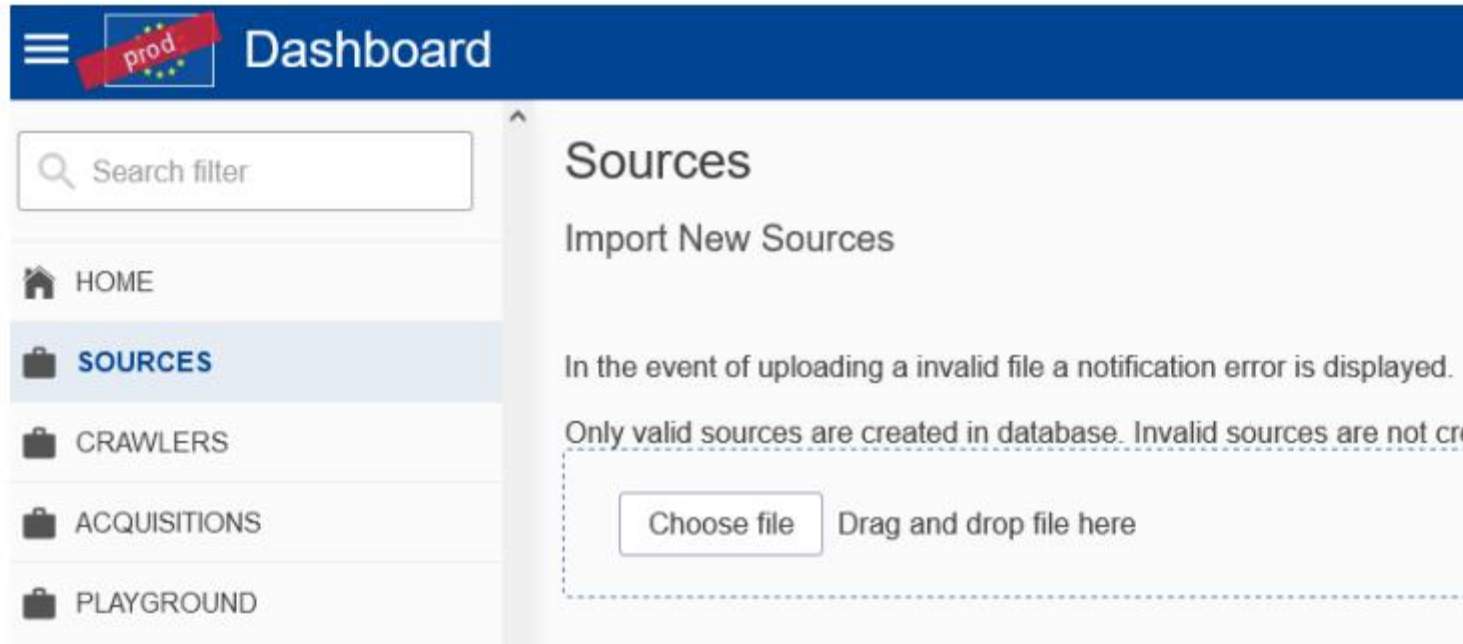
4. Opening WIP – cloning the Github open code on SMP.

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.



2. Uploading the dataset of URLs into WIP.



The screenshot shows the WIP Dashboard interface. The top navigation bar is dark blue with a hamburger menu icon, a 'prod' badge, and the word 'Dashboard'. Below the navigation bar is a search filter input field. The left sidebar contains a list of menu items: HOME, SOURCES (highlighted), CRAWLERS, ACQUISITIONS, and PLAYGROUND. The main content area is titled 'Sources' and features an 'Import New Sources' section. This section includes a warning message: 'In the event of uploading a invalid file a notification error is displayed. Only valid sources are created in database. Invalid sources are not cre'. Below the message is a dashed box containing a 'Choose file' button and the text 'Drag and drop file here'.



Collecting and processing data in 6 steps

1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

4. Opening WIP – cloning the Github open code on SMP.

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.



3. Modifying necessary parameters and starting the Crawler.

Create a new crawler [Help](#)

Crawler

Name *

Group *

Data Acquisition

Create a data acquisition

Crawler Name *

Workflow ID *



Question?

- Which solution do you prefer for webscraping:
 - a) centralized one with big scalable storage in NoSQL and SQL database maintained by Eurostat, access to Python JupyterLab, Rstudio
 - b) local one in-house in your institute



Web Intelligence
Network



Funded by
the European Union

Collecting and processing data in 6 steps

1. Starting with showing the dataset of 5 enterprises with anonymized Business Register numbers.

2. Uploading the dataset of URLs into WIP

3. Modifying necessary parameters and starting the Crawler.

4. Opening Datalab – cloning the Github open code on SMP.

5. Executing the software in Python and accessing the data from Crawler already scrapped.

6. Waiting for the results and open CSV file already prepared.

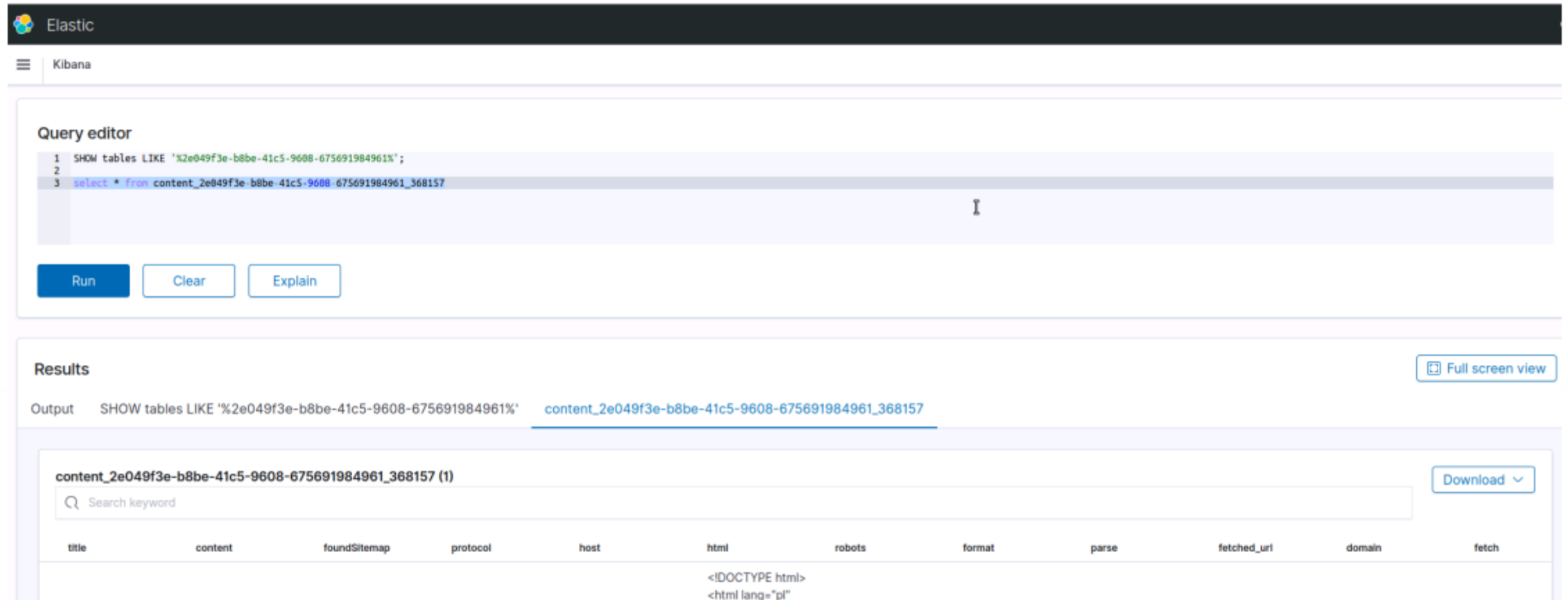


4. Opening WIP – cloning the Github open code on SMP.

```
jmalankowski@ip-10-0-115-136:~$ git clone https://github.com/jmaslankowski/StarterKit
Cloning into 'StarterKit'...
remote: Enumerating objects: 243, done.
remote: Total 243 (delta 0), reused 0 (delta 0), pack-reused 243
Receiving objects: 100% (243/243), 925.87 KiB | 3.40 MiB/s, done.
Resolving deltas: 100% (129/129), done.
jmalankowski@ip-10-0-115-136:~$
```



Showing the existence of the data in NoSQL



The screenshot shows the Elastic Kibana interface. At the top, the Elastic logo and 'Kibana' are visible. Below is the 'Query editor' section with a text area containing a SQL query:

```
1 SHOW tables LIKE '%2e049f3e-b8be-41c5-9608-675691984961%';  
2  
3 select * from content_2e049f3e-b8be-41c5-9608-675691984961_368157
```

Below the query editor are three buttons: 'Run', 'Clear', and 'Explain'. The 'Results' section below shows the output of the query:

Output SHOW tables LIKE '%2e049f3e-b8be-41c5-9608-675691984961%' [content_2e049f3e-b8be-41c5-9608-675691984961_368157](#)

Below the output is a search bar and a 'Download' button. The search results are displayed in a table with the following columns:

title	content	foundSitemap	protocol	host	html	robots	format	parse	fetchesd_url	domain	fetch
					<!DOCTYPE html> <html lang="pl"						

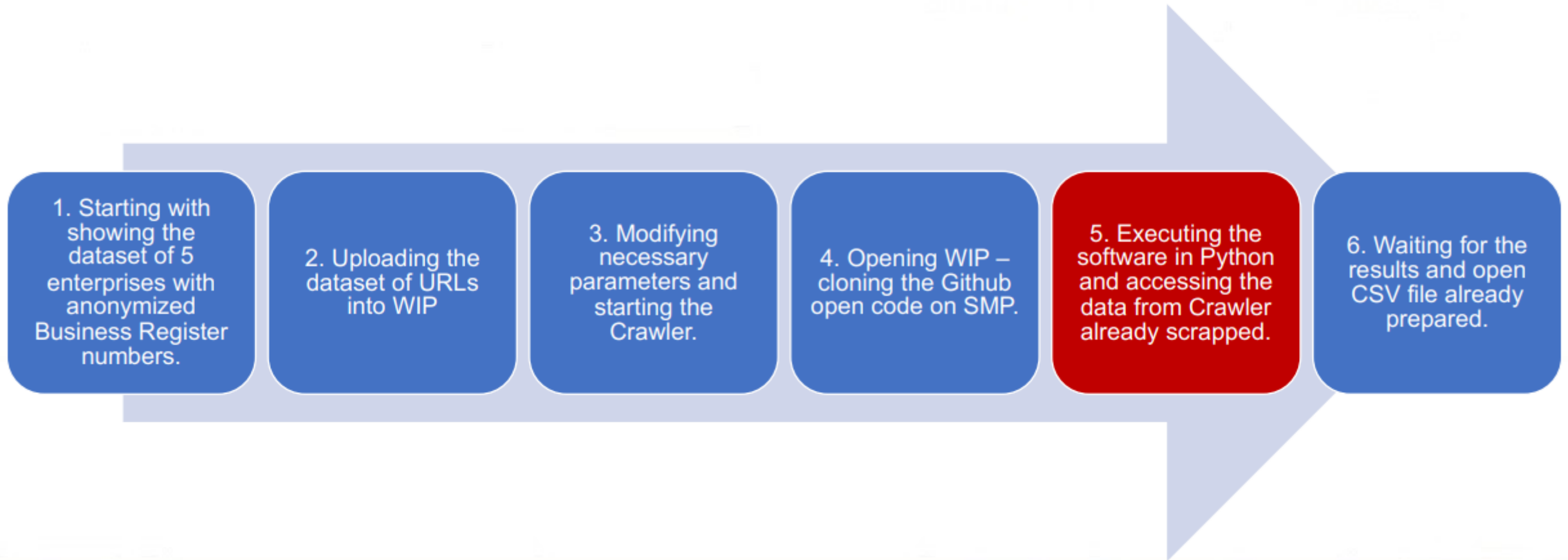


Question?

- What is more convenient for you, store the webpages in:
 - a) NoSQL
 - b) flat files
 - c) SQL
 - d) do not store webpages but processing necessary attributes in-memory



Collecting and processing data in 6 steps



5. Executing the software in Python and accessing the data from Crawler already scrapped.

```
In [13]: body={
  "query": {
    "query_string": {
      "query": "stat.gov.pl"
    }
  }
}

resp = elastic.search(index="content_*", body=body)
print("Got %d Hits:" % resp['hits']['total']['value'])
htmlfile=""
for doc in resp['hits']['hits']:
    #print("%s\n %s" % (doc['_id'], doc['_source']))
    htmlfile=str(doc['_source']['html'])
    url=str(doc['_source']['fetch_url'])
    try:
        date_1=str(doc['_source']['protocol.last-modified'])
    except:
        print("ERROR")
    print("",url,date_1)

print(doc['_source'].keys())
print(doc.keys())
```

Got 8 Hits:

```
https://stat.gov.pl Wed, 23 Nov 2022 11:05:31 GMT
https://stat.gov.pl Mon, 16 Jan 2023 07:06:58 GMT
https://stat.gov.pl Mon, 30 Jan 2023 11:11:34 GMT
https://stat.gov.pl Tue, 31 Jan 2023 11:24:55 GMT
https://stat.gov.pl Mon, 06 Feb 2023 14:01:00 GMT
https://stat.gov.pl Mon, 31 Oct 2022 01:16:30 GMT
https://stat.gov.pl Tue, 31 Jan 2023 11:25:30 GMT
https://stat.gov.pl Mon, 06 Feb 2023 14:01:36 GMT
```

```
In [12]: import WPCStarterKit as wsk
smp=wsk.SocialMediaPresence()
smp.searchSocialMediaLinks(url,htmlfile)
```

The length of the scrapped content: 92415 characters

Number of links on website: 259

<https://www.youtube.com/channel/UC0wiQME1FgYszpAoYgTnXtg/featured>

<https://www.facebook.com/GlownyUrzadStatystyczny/>

http://twitter.com/GUS_STAT

<https://www.linkedin.com/company/532930>

https://www.instagram.com/gus_stat/

https://twitter.com/GUS_STAT/lists/gus-i-urz-dy-statystyczne?ref_src=twsrc%5Etfw

Total number of unique social media links found: 6

```
Out[12]: {'URL': 'http://stat.gov.pl',
'Facebook': ['https://www.facebook.com/GlownyUrzadStatystyczny/'],
'Twitter': ['https://twitter.com/GUS_STAT/lists/gus-i-urz-dy-statystyczne?ref_src=twsrc%5Etfw',
'http://twitter.com/GUS_STAT'],
'Youtube': ['https://www.youtube.com/channel/UC0wiQME1FgYszpAoYgTnXtg/featured'],
'LinkedIn': ['https://www.linkedin.com/company/532930'],
'Instagram': ['https://www.instagram.com/gus_stat/'],
'Xing': [],
'Pinterest': []}
```

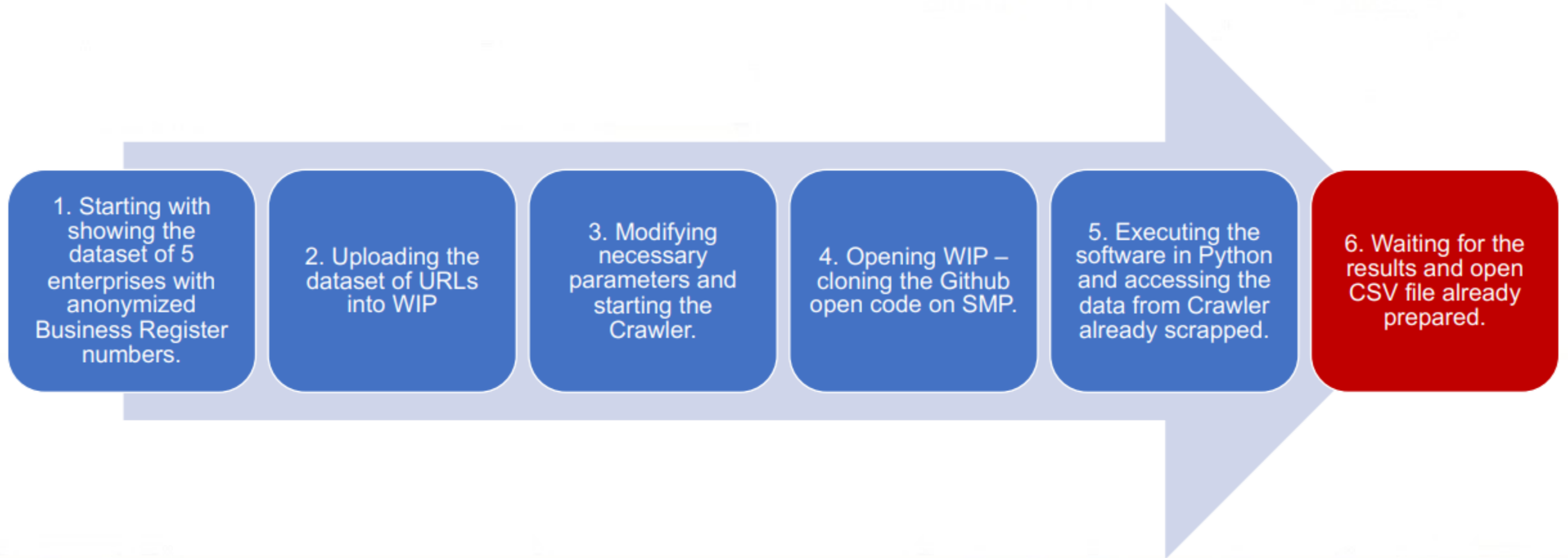


Question?

- What language do you prefer the most (in alphabetical order):
 - A) Java
 - B) Python
 - C) R
 - D) do not know any of them



Collecting and processing data in 6 steps



6. Waiting for the results and open CSV file already prepared.

jupyter wp2_social.csv ✓ kilka sekund temu

File Edit View Language current mode

```
1 URL;Facebook;Twitter;Youtube;LinkedIn;Instagram;Xing;Pinterest
2 http://stat.gov.pl;https://www.facebook.com/GlownyUrzadStatystyczny/;https://twitter.com/GUS_STAT/lists/gus-i-urz-dy-
  statystyczne?ref_src=twsrc%5Etfw ,http://twitter.com/GUS_STAT;https://www.youtube.com/channel/UC0wiQMElFgYszpAoYgTnXtg/featured;https:
  //www.linkedin.com/company/532930;https://www.instagram.com/gus_stat/;
```

	A	B	C	D	E	F	
URL	Facebook	Twitter	Youtube	LinkedIn	Instagram		Xi
	http://stat.gov.pl	https://www.facebook.com/GlownyUrzadStatystyczny/	https://twitter.com/GUS_STAT	https://www.youtube.com/channel/UC0wiQMElFgYszpAoYgTnXtg/	https://www.linkedin.com/company/532930	https://www.instagram.com/gus_stat/	



Web Intelligence
Network



Funded by
the European Union

Thank you!
Jacek Maślankowski
j.maslankowski@stat.gov.pl



Web Intelligence
Network



Funded by
the European Union