WIH platform: Architectural overview

Mauro Bruno (ISTAT)
Giuseppina Ruocco (ISTAT)

23 November 2022

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence
Network

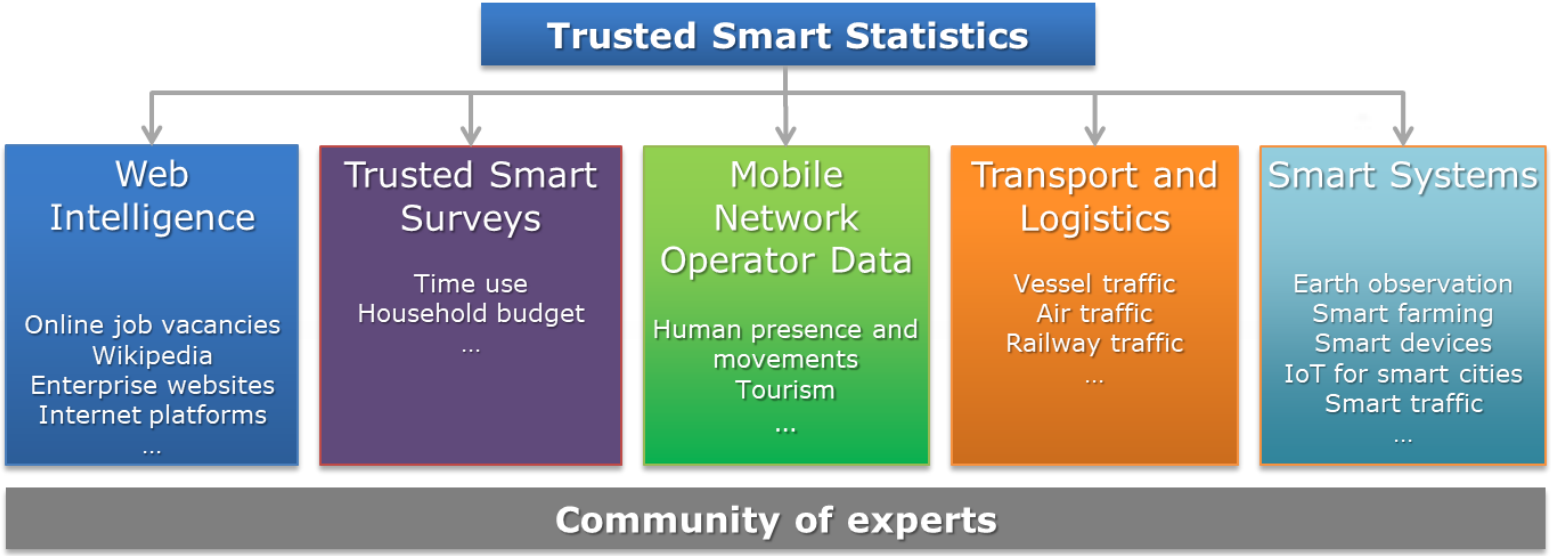
Funded by
the European Union

Outline

- Main goals of the Web Intelligence Network (WIN)
- Web Intelligence Hub (WIH) services
- Web Intelligence Hub (WIH) architecture
- BREAL and WIH implementation
- WIH services: **NSI’s perspective**

Web Intelligence Network (WIN)

- Support National Statistical Institutes (NSIs) in the use of tools and technologies for **web data collection** and processing (such as Web Scraping, Natural Language Processing, Machine Learning)
- Develop, test and document **reusable tools** for collecting and processing web data



Trusted Smart Statistics

Web Intelligence
Online job vacancies
Wikipedia
Enterprise websites
Internet platforms
…

Trusted Smart Surveys
Time use
Household budget
…

Mobile Network Operator Data
Human presence and movements
Tourism
…

Transport and Logistics
Vessel traffic
Air traffic
Railway traffic
…

Smart Systems
Earth observation
Smart farming
Smart devices
IoT for smart cities
Smart traffic
…

Community of experts

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

Funded by
the European Union

Web Intelligence Network (WIN)

- Support National Statistical Institutes (NSIs) in the use of tools and technologies for **web data collection** and processing (such as Web Scraping, Natural Language Processing, Machine Learning)
- Develop, test and document **reusable tools** for collecting and processing web data

Web Intelligence

Online job vacancies
Wikipedia
Enterprise websites
Internet platforms
…

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

Funded by the European Union

Web Intelligence Hub (WIH) services to support ESS

- Data acquisition (web scraping, 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 data

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

Funded by the European Union

Web Intelligence Hub (WIH) services to support ESS

- Data acquisition (web scraping, 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

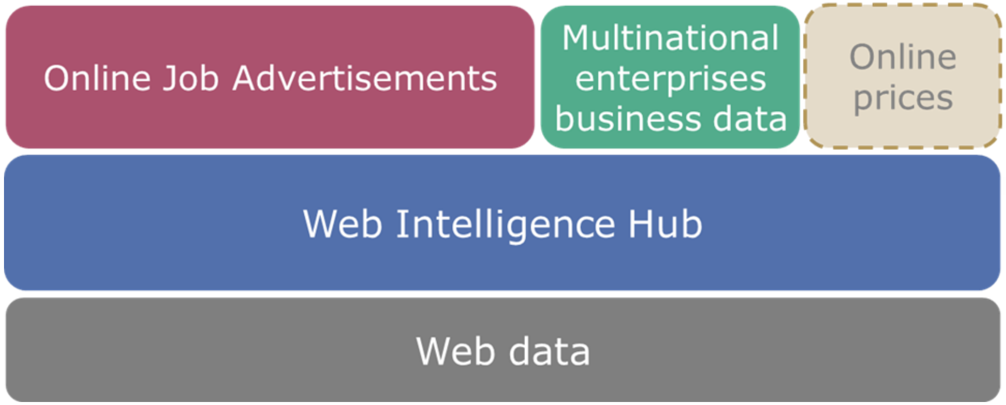
Web Intelligence Hub
Web data

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

Funded by the European Union

Web Intelligence Hub (WIH) services to support ESS

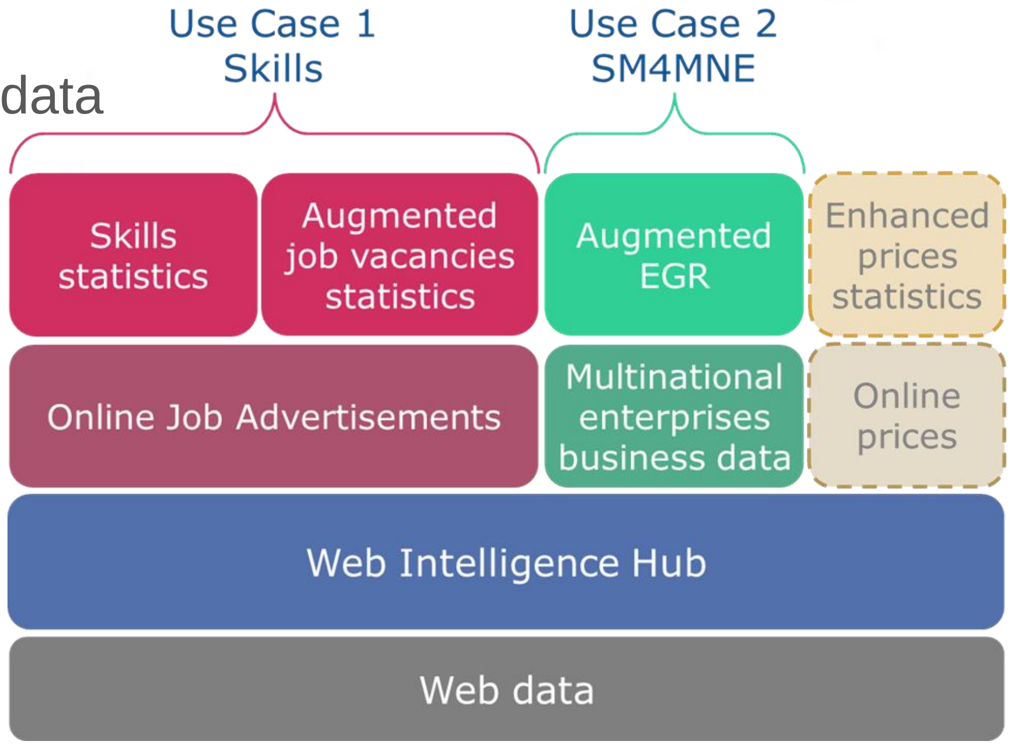
- Data acquisition (web scraping, APIs)
- Trans-national data agreements
- Partnership models for national data agreements
- **IT infrastructure and tools**
- Analytical services (e.g., NLP)
- **Methodology**
- Regulatory aspects
- **Skills (training material)**
- **R&D collaboration**
- Governance


Online Job Advertisements
Multinational enterprises business data
Online prices
Web Intelligence Hub
Web data

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

Web Intelligence Hub (WIH) services to support ESS

- Data acquisition (web scraping, 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



Use Case 1
Skills
Use Case 2
SM4MNE
Skills statistics
Augmented job vacancies statistics
Augmented EGR
Enhanced prices statistics
Online Job Advertisements
Multinational enterprises business data
Online prices
Web Intelligence Hub
Web data

Source: Trusted Smart Statistics Centre Web Intelligence Hub
https://ec.europa.eu/eurostat/cros/content/11-web-intelligence-hub-presentation_en

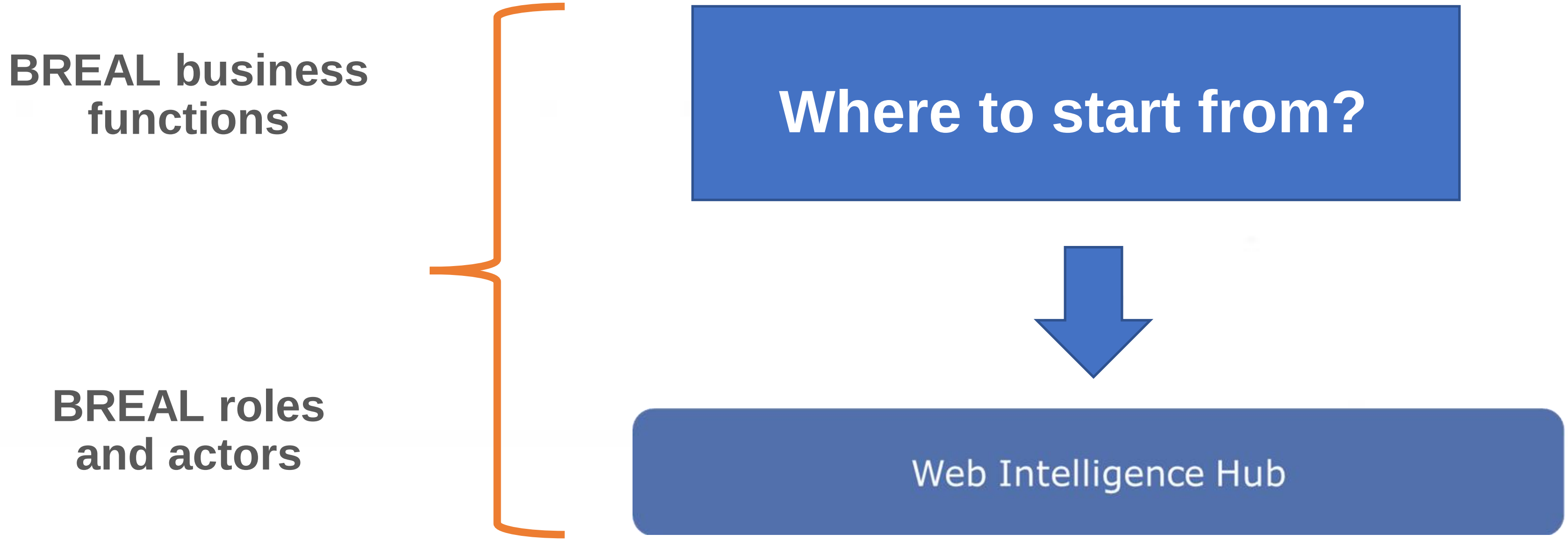
Funded by the European Union

Web Intelligence Hub (WIH) architecture

Web Intelligence Hub

Funded by
the European Union

Web Intelligence Hub (WIH) architecture



BREAL business functions

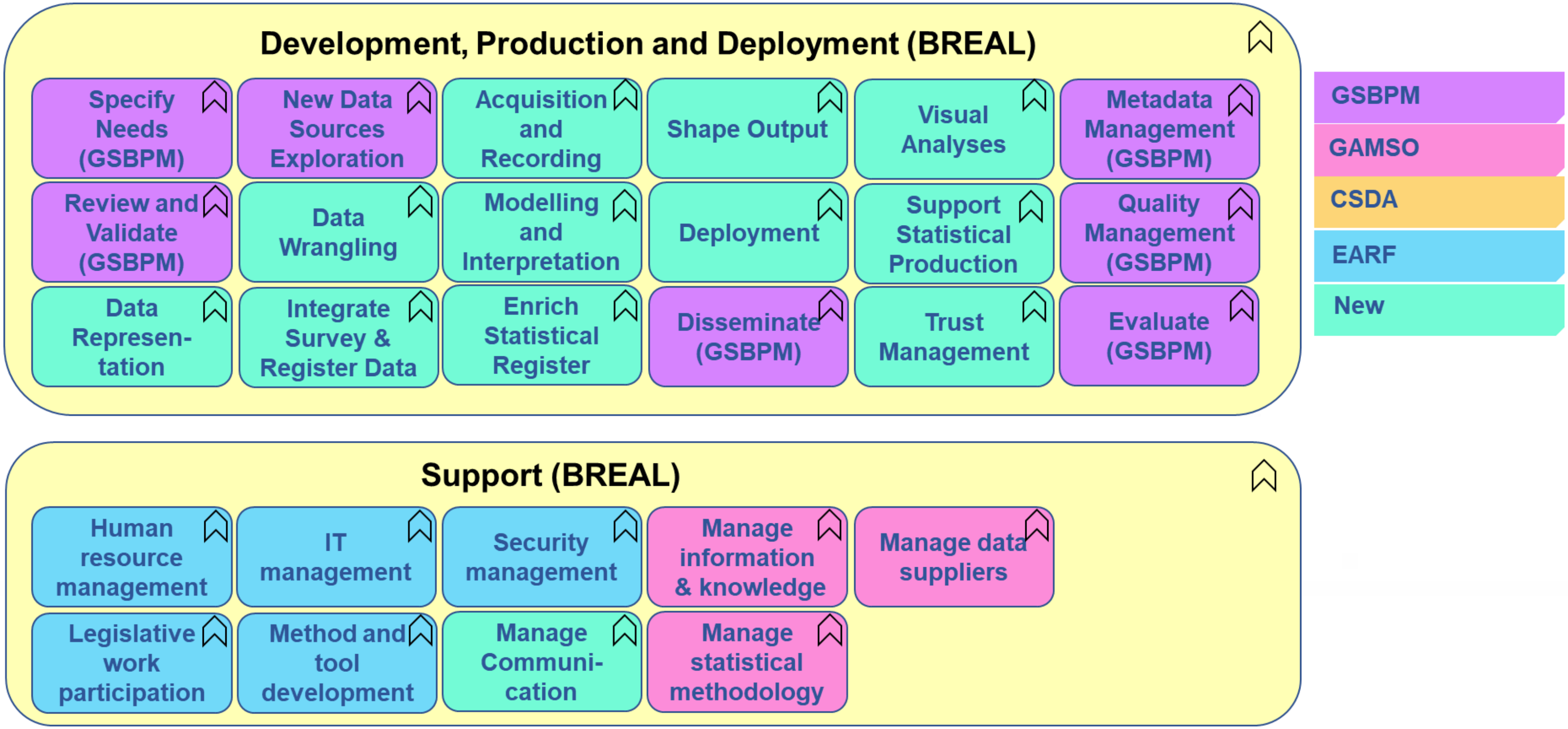
BREAL roles and actors

Where to start from?

Web Intelligence Hub

Funded by the European Union

BREAL and WIH implementation (1)



Development, Production and Deployment (BREAL)

Specify Needs (GSBPM)	New Data Sources Exploration	Acquisition and Recording	Shape Output	Visual Analyses	Metadata Management (GSBPM)
Review and Validate (GSBPM)	Data Wrangling	Modelling and Interpretation	Deployment	Support Statistical Production	Quality Management (GSBPM)
Data Represen-tation	Integrate Survey & Register Data	Enrich Statistical Register	Disseminate (GSBPM)	Trust Management	Evaluate (GSBPM)

Support (BREAL)

Human resource management	IT management	Security management	Manage information & knowledge	Manage data suppliers
Legislative work participation	Method and tool development	Manage Communi-cation	Manage statistical methodology	

GSBPM

GAMSO

CSDA

EARF

New

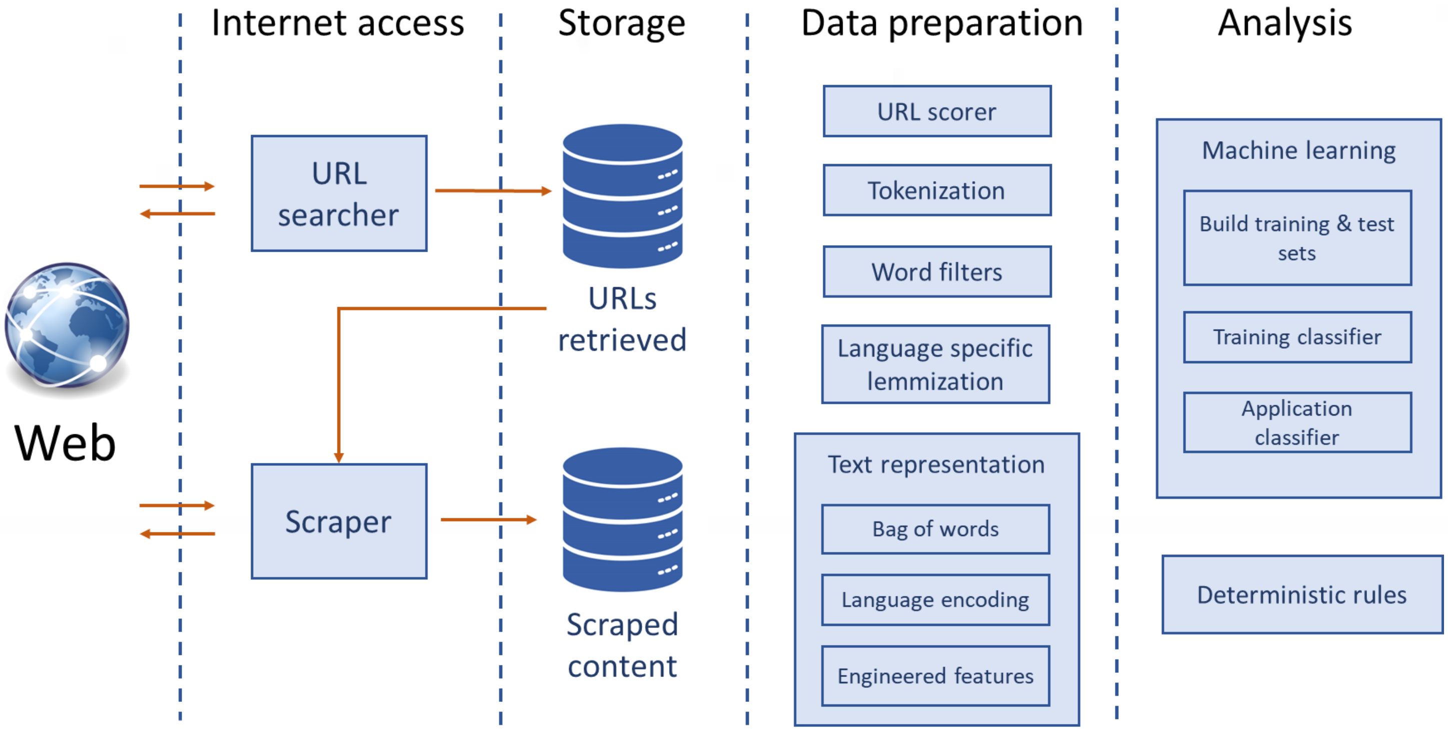
Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer B., Kostadin G., Paulussen R., Quaresma S. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT

Web Intelligence Network

Funded by the European Union

BREAL and WIH implementation (2)

BREAL business functions and **OBEC** implemented workflow



Internet access

Storage

Data preparation

Analysis

Web

URL searcher

Scraper

URLs retrieved

Scraped content

URL scorer

Tokenization

Word filters

Language specific lemmization

Text representation

Bag of words

Language encoding

Engineered features

Machine learning

Build training & test sets

Training classifier

Application classifier

Deterministic rules

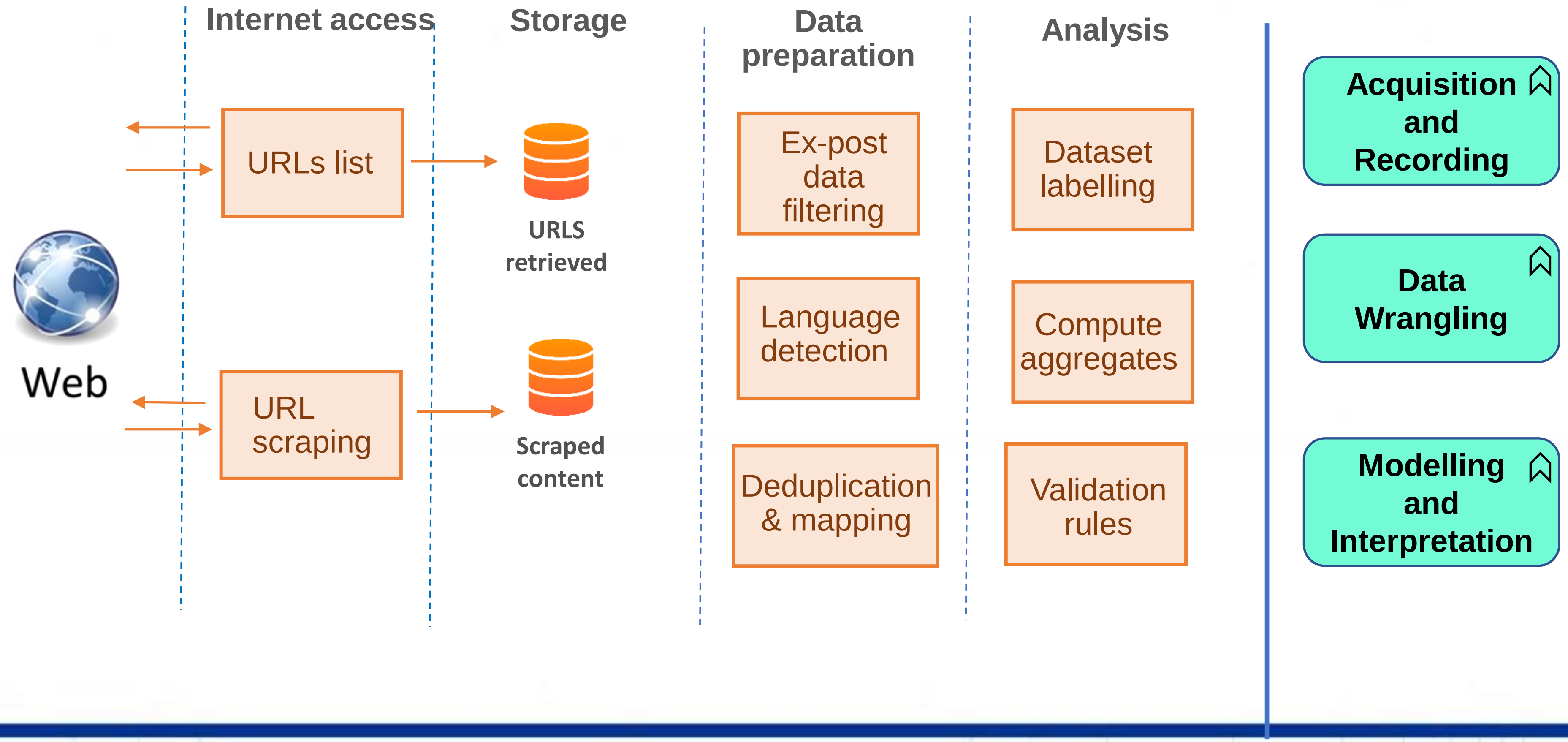
Acquisition and Recording

Data Wrangling

Modelling and Interpretation

BREAL and WIH implementation (3)

BREAL business functions and **OJA** implemented workflow



Internet access

Storage

Data preparation

Analysis

Web

URLs list

URL scraping

URLS retrieved

Scraped content

Ex-post data filtering

Language detection

Deduplication & mapping

Dataset labelling

Compute aggregates

Validation rules

Acquisition and Recording

Data Wrangling

Modelling and Interpretation

Funded by the European Union

WIH services: NSI’s perspective (1)

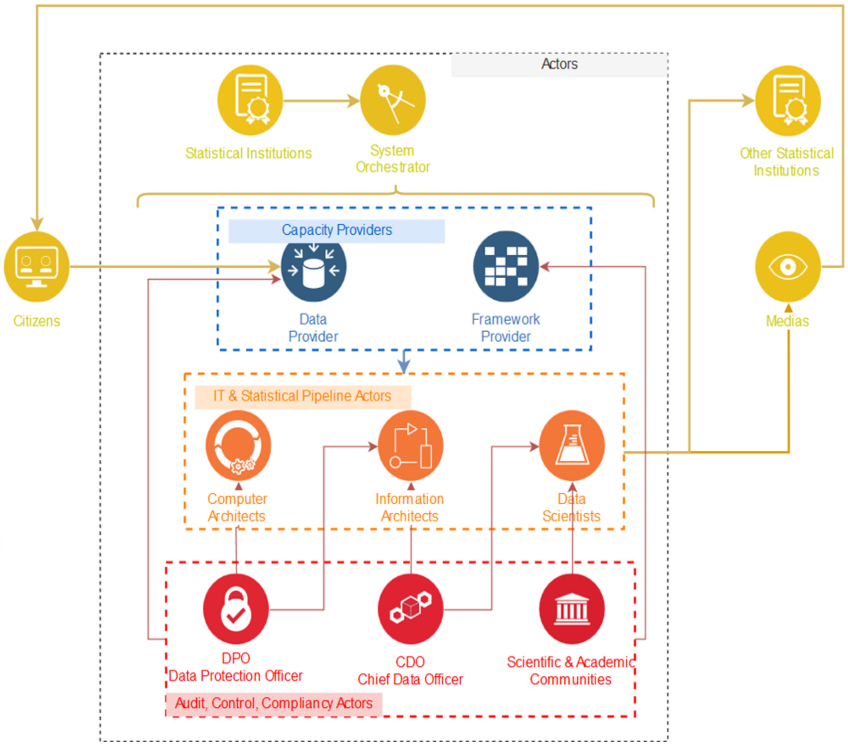
NSI’s staff accessing the WIH

Data scientist

Researcher

Domain specialist

BREAL actors and stakeholders



Source: Scannapieco M.,. et al. (2019): BREAL. Big Data Reference Architectu and Layers. Version 2019-12-09. Edited by EUROSTAT

WIH services: NSI’s perspective (1)

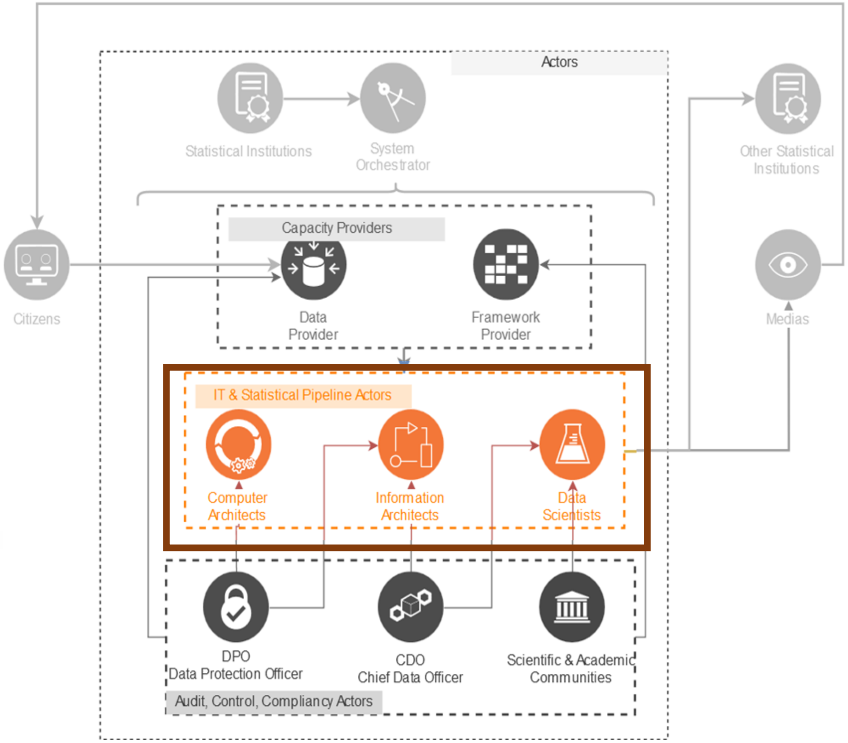
NSI’s staff accessing the WIH

Data scientist

Researcher

Domain specialist

BREAL actors and stakeholders



Source: Scannapieco M.,. et al. (2019): BREAL. Big Data Reference Architectu and Layers. Version 2019-12-09. Edited by EUROSTAT

WIH services: the NSI’s perspective (2)

How to approach the BREAL functions?

Through **use cases** to analyze what could be **reused** and **standardized**

How to approach Actors/Roles?

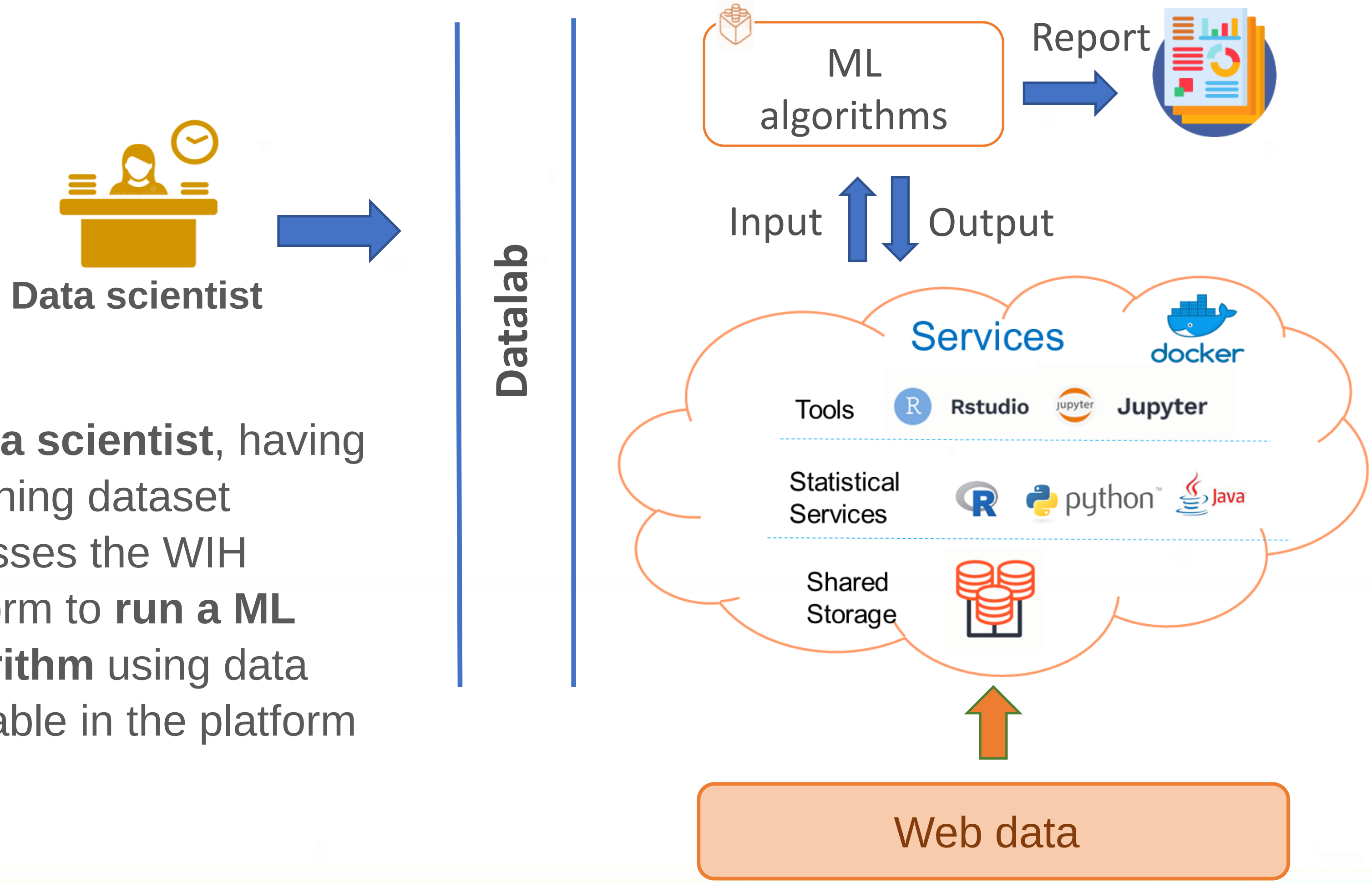
Through **user stories** to highlight NSI’s perspective

User stories

1. **Using big data capabilities**
2. **Harmonizing traditional and big data sources**
3. **Analyzing statistical output from web data**

Funded by the European Union

1. Using big data capabilities

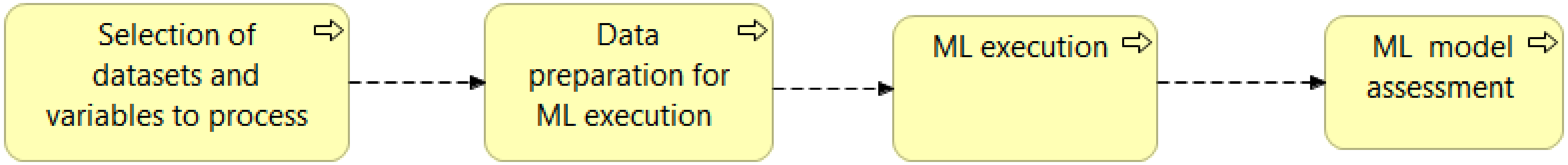


A **data scientist**, having a training dataset accesses the WIH platform to **run a ML algorithm** using data available in the platform

Funded by the European Union

1. Using big data capabilities: the data scientist perspective

Tasks to execute


Selection of datasets and variables to process
Data preparation for ML execution
ML execution
ML model assessment

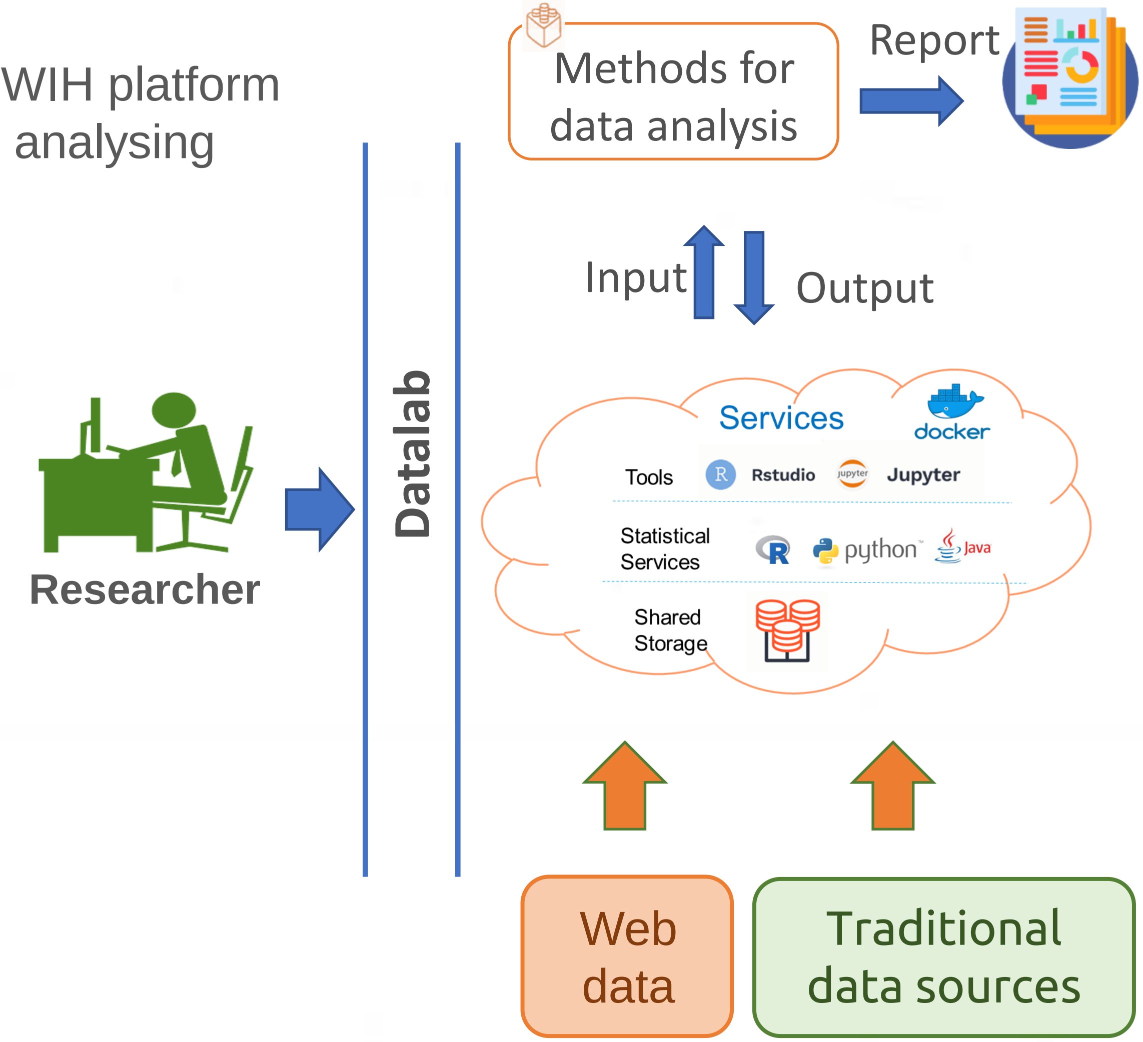
Data and metadata management

- Data structure description to select the datasets and variables of interest
- Tracking of the main process steps for process auditability and reproducibility
- ML quality indicators

2. Harmonizing traditional and big data sources

A **researcher** may access the WIH platform to **run statistical methods** for analysing web data to:

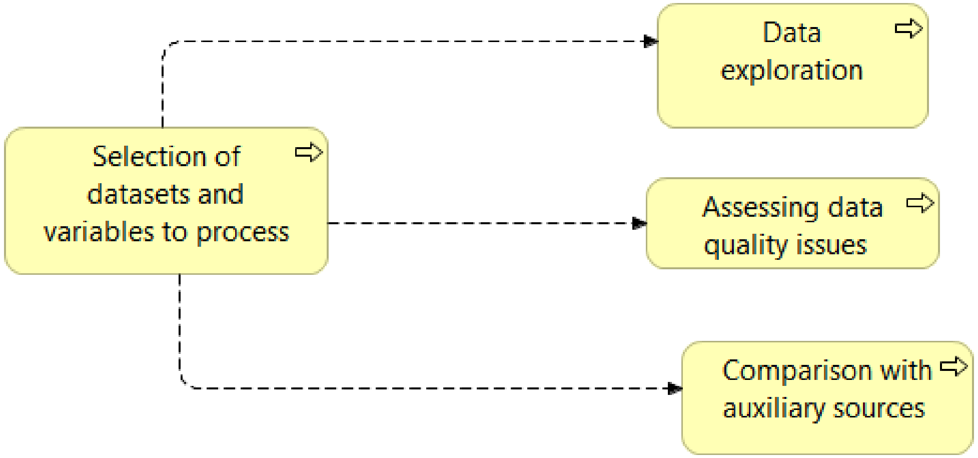
- Enrich or reduce the amount of information collected through traditional survey modes
- Test different methods of record linkage, to combine survey and web data sources
- Provide an assessment of web data sources in terms of representativeness of the statistical population
- Highlight coverage issues affecting specific subsets of units…



Funded by the European Union

2. Harmonizing traditional and big data sources: the researcher perspective

Tasks to execute



Selection of datasets and variables to process

Data exploration

Assessing data quality issues

Comparison with auxiliary sources

Data and metadata management

- Description of data structures to select the datasets and variables of interest
- Tracking of the main process steps for process auditability and reproducibility
- Indicators for assessing the output of applied methods

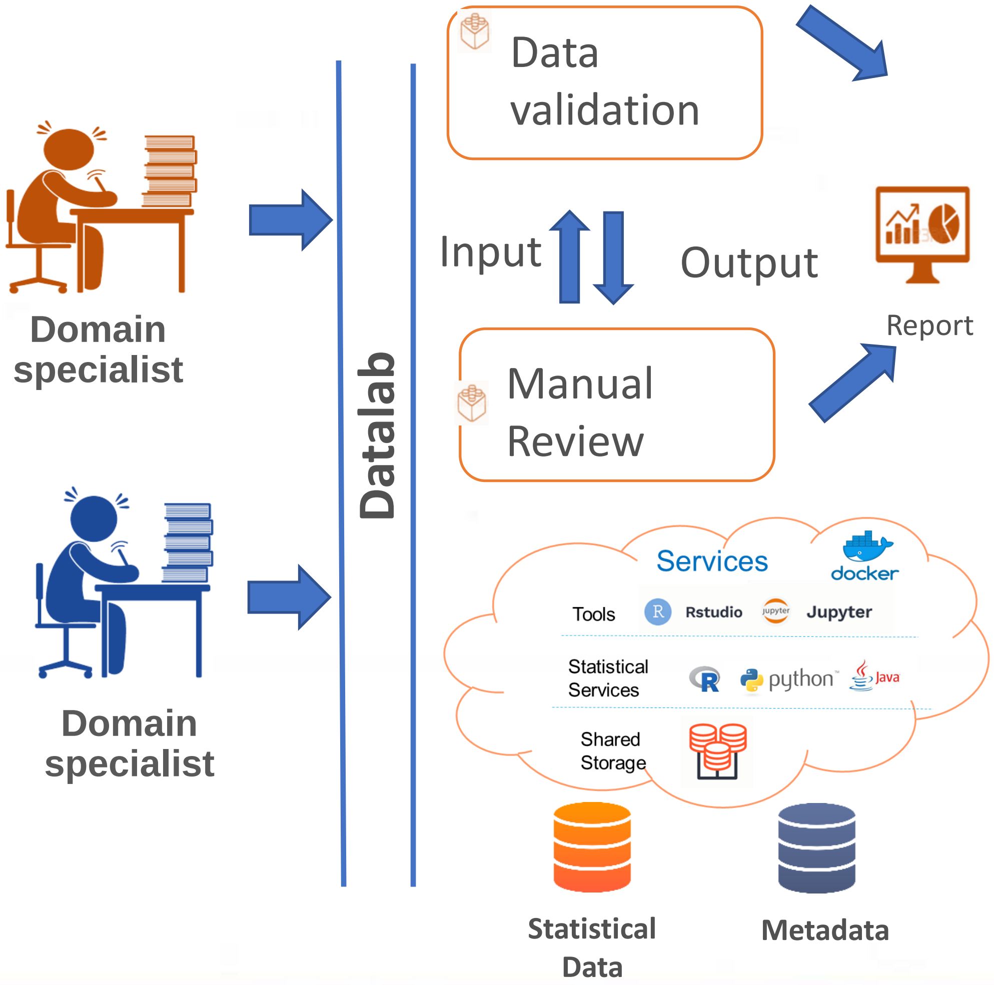
Web Intelligence Network

Funded by the European Union

3. Analyzing statistical output from web data

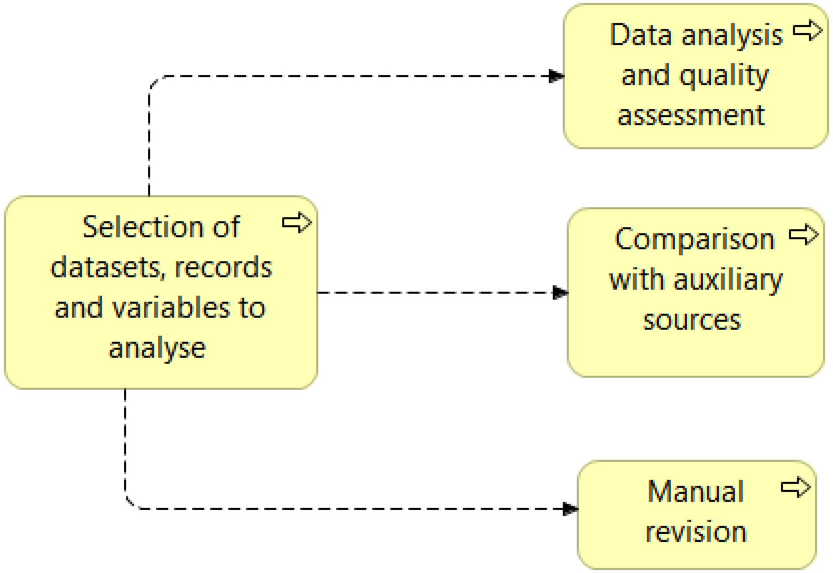
Domain specialists, involved in the statistical production may access the WIH platform to contribute to the data validation process through:

- A benchmark of aggregated statistical output extracted from web data with auxiliary data sources and official statistics
- Assessment of data accuracy in terms of coherence and comparability
- Manual revision of statistical output to validate and improve the WIH data workflow



3. Analyzing statistical output from web data: the domain specialist perspective

Tasks to execute


Selection of datasets, records and variables to analyse

Data analysis and quality assessment

Comparison with auxiliary sources

Manual revision

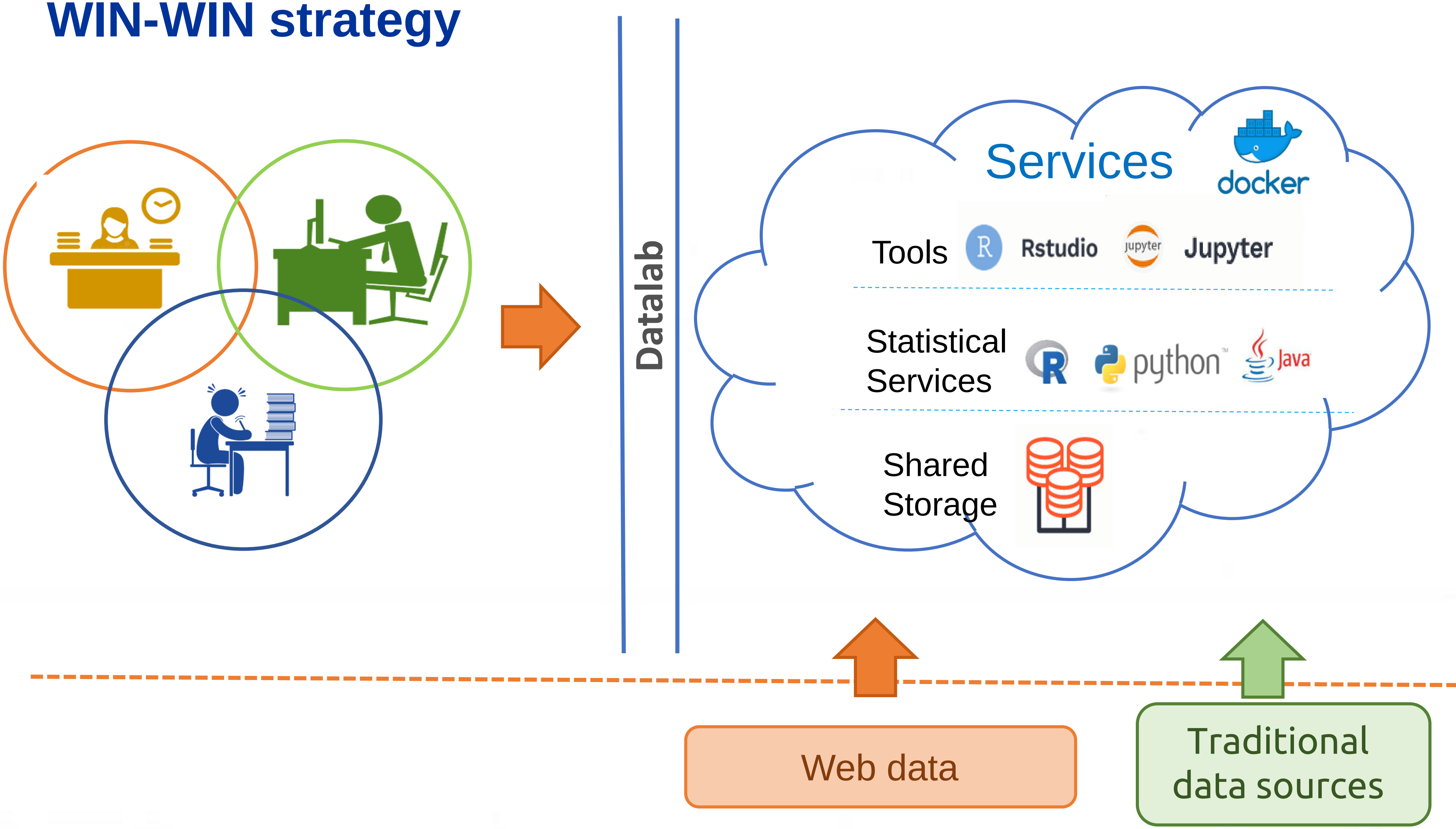
Data and metadata management

- Description of data structures to select the datasets and variables of interest
- Tracking of the main process steps for process auditability and reproducibility
- Indicators for assessing the output of applied methods

Web Intelligence Network

Funded by the European Union

WIN-WIN strategy



Datalab

Services

docker

Tools

Rstudio

Jupyter

Statistical Services

python

Java

Shared Storage

Web data

Traditional data sources

Thank you for your attention!

giruocco@istat.it

mbruno@istat.it

WIN, WP4
Overview of Quality Aspects

Alexander Kowarik (Statistik Austria)
Magdalena Six (Statistik Austria)
23 November

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

Outline

- Quality of traditional data sources
  vs.
  Quality of new data sources / Quality of web data

- Examples of quality aspects w.r.t the data pipeline of OJA
- Quality guidelines for web data (Deliverable 4.1)

- Example of stable data access
- Example of coverage
- Example of comparability over time

Quality in case of **traditional data sources** (samples, admin data)

Quality dimensions and quality indicators: well-known, widely accepted, part of ESS-wide quality reporting (eg. SIMSv2*, EHQMR)

- **Relevance** (S12)
- **Accuracy** and **Reliability** (S13)
  e.g. Unit Non Response, Item Non Response Rate (A4, A5),
  Over-Coverage Rate (A2)..)
- **Timeliness** and **Punctuality** (S14)
- **Comparability** and **Coherence** (S15)
  e.g. Lenght of comparable time series (CC2)
- **Accessability** and **Clarity** (S10)

*Overview over SIMSv2:
https://ec.europa.eu/eurostat/documents/64157/4373903/SIMS-2-0-Revised-standards-November-2015-ESSC-final.pdf

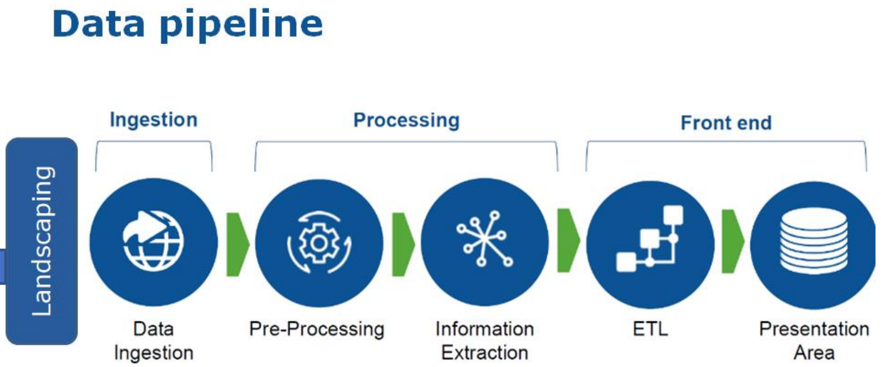
Quality of **new data sources**

- Traditional quality dimensions **not sufficient** to cover new aspects in production line, **new quality indicators** needed/wanted!
- Focus on **access to data** and part of the **throughput phase**
  (BREAL functions «acquisition and recording», «data wrangling», «pre-processing»)
- Traditional quality dimensions and quality indicators sometimes applicable (e.g. overcoverage rate), sometimes not meaningul (e.g. response rates)
- **New, not yet well established indicators** necessary:
  e.g. duration of stable access, duplication rate in case of OJA
- Since new data sources vary hugely, quality considerations should take data classes into account, e.g. Data class «web data»

Funded by
the European Union

Examples for quality apects and data pipeline

Data Pipeline taken from WIH OJA data



Which websites w.r.t the wanted information **exist**? (job portals, real estate web portals…)

Which websites can be **accessed** and how (API, sraped)?

Which websites promise **stable access**?

Which **variables** are available (on all websites)?

Monitor & control: Data ingestion performances are affected by external elements. Data ingestion processes can fail for different causes (e.g. loss of connectivity; internal server errors from job portals; websites changes)

Merging different data sources

De-Duplication: Three parts (for OJA)
- Physical deduplication
- Semantic deduplication
- Logic deduplication

Coverage & Selectivity Penetration of OJV markets varies in and across countries and may change over time

Model & Processing Errors Various error sources when using ontologies, text mining, machine learning methods…

Data Accuracy: miss-classification may be encountered regarding the occupation (ISCO), skills (ISCO), geo localisation (NUTS), economic activities (NACE)

Web Intelligence Network

Funded by the European Union

Deliverable 4.1: Minimal guidelines and recommendations for implementation (quality part)

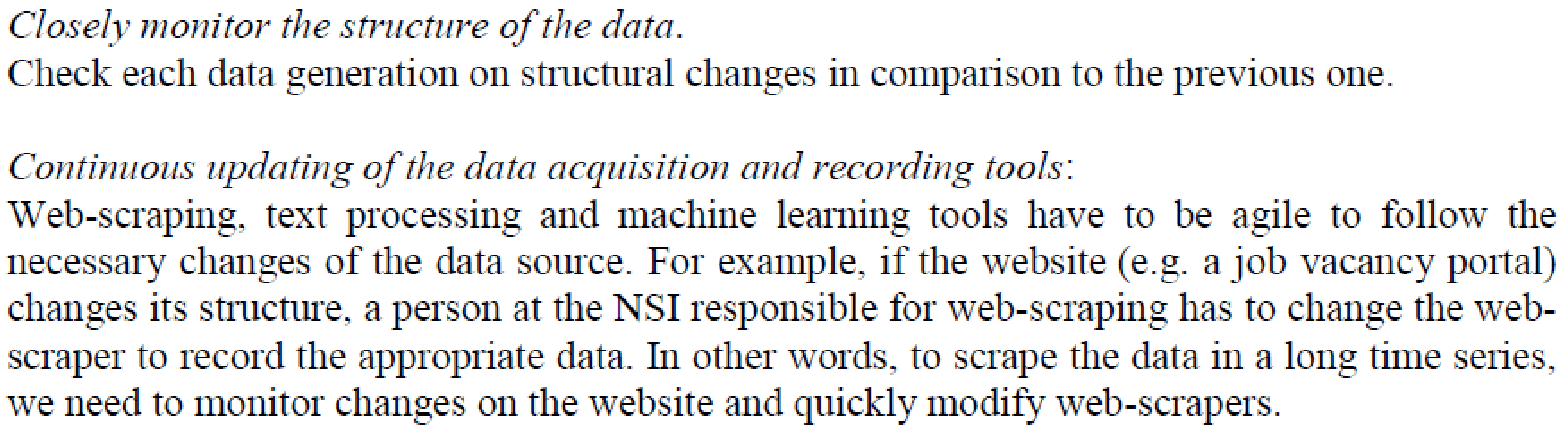
E.g. for the throughput phase, guidelines for the following quality aspects are listed:

- **Linking**
- **Coverage**
- **Comparability over time**
- **Measurement errors**
- **Model errors / Process errors**

Guidelines for Web data: https://ec.europa.eu/eurostat/cros/content/deliverable-41-minimal-guidelines-and-recommendations-implementation_en
General guidelines for new data sources:
https://ec.europa.eu/eurostat/cros/sites/default/files/WP3_Deliverable_K3_Revised_Version_of_the_Quality_Guidelines_for_the_Acquisition_and_Usage_of_Big_Data_Final_version.pdf

Deliverable 4.1: Minimal guidelines and recommendations for implementation (quality part)

Example of Guidelines - **Comparability over time**



Closely monitor the structure of the data.
Check each data generation on structural changes in comparison to the previous one.

Continuous updating of the data acquisition and recording tools:
Web-scraping, text processing and machine learning tools have to be agile to follow the necessary changes of the data source. For example, if the website (e.g. a job vacancy portal) changes its structure, a person at the NSI responsible for web-scraping has to change the web-scraper to record the appropriate data. In other words, to scrape the data in a long time series, we need to monitor changes on the website and quickly modify web-scrapers.

Minimal Guidelines for Web data: https://ec.europa.eu/eurostat/cros/content/deliverable-41-minimal-guidelines-and-recommendations-implementation_en

Example: Data Access
Which websites to scrape? What to scrape?

WP3 agreed for all Use Cases on a list of common criteria*:

Mandatory variables are a set of variables whose presence in the advertisements, published on the web data sources, is compulsory and can be extracted from any ad for all selected sources for a given use case.

Checklist for quality assessment of web data sources
The sources should be ranked based on a score (range 0-100), where a score of 0 indicates that the source is rejected.

Stop Criteria - If at least one Stop criteria has a value of 1, then the web source is rejected. Usually, it happens when captcha or robot blocking mechanisms were built on the website

…

*All information from the first interim report (Deliverable 3_1, not public), general information about WP3 available here: https://ec.europa.eu/eurostat/cros/content/work-package-3-%E2%80%93-new-use-cases_en

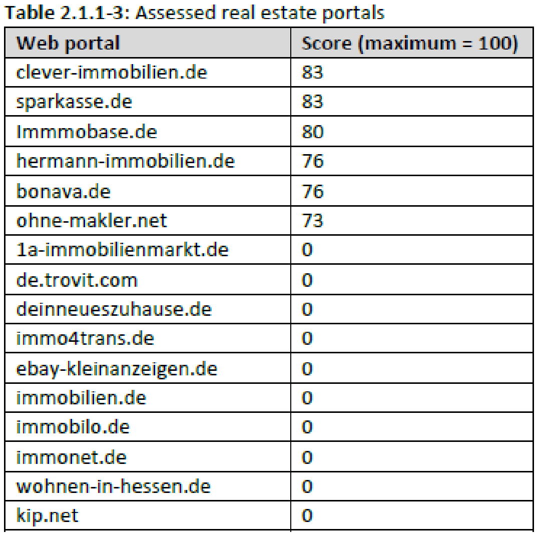


Table 2.1.1-3: Assessed real estate portals

Web portal	Score (maximum = 100)
clever-immobilien.de	83
sparkasse.de	83
Immmobase.de	80
hermann-immobilien.de	76
bonava.de	76
ohne-makler.net	73
1a-immobilienmarkt.de	0
de.trovit.com	0
deinneueszuhause.de	0
immo4trans.de	0
ebay-kleinanzeigen.de	0
immobilien.de	0
immobilo.de	0
immonet.de	0
wohnen-in-hessen.de	0
kip.net	0

Table from Del.3_1_WP3 *, UC1, Example for assessed and rated real-estate portals for Germany

Example: Duplicates

E.g. UC1 “Characteristics of the real estate market”, France

*“Moreover, one of the main challenges one must deal with regarding the data is that, although the advertisements are uniquely identified, they are **not unique**, in the sense **that several ads may refer to the same offer and the same dwelling.** Methodological analyses on the text have been undertaken to **identify and remove duplicates**.“*

Funded by the European Union

Example: Duplicates

E.g. WP2 / OJA, on the basis of CEDEFOP data:

Paper: De Lazzer & Rengers (2021), «Impact of the corona virus on the labour market: experimental statistics based on data from online job portals»

“Employers often post vacancies on several job boards simultaneously, thus making it harder to assess the actual stock of jobs that are available. In its automatic processing of data, CEDEFOP classifies a proportion of the job advertisements as duplicates, with the most ***important criteria being whether their content is similar*** *and whether they were* ***published at around the same time****. However, there is no empirical way of verifying how successful this process of deduplication is, and CEDEFOP was also unable to gauge its reliability when asked.”*

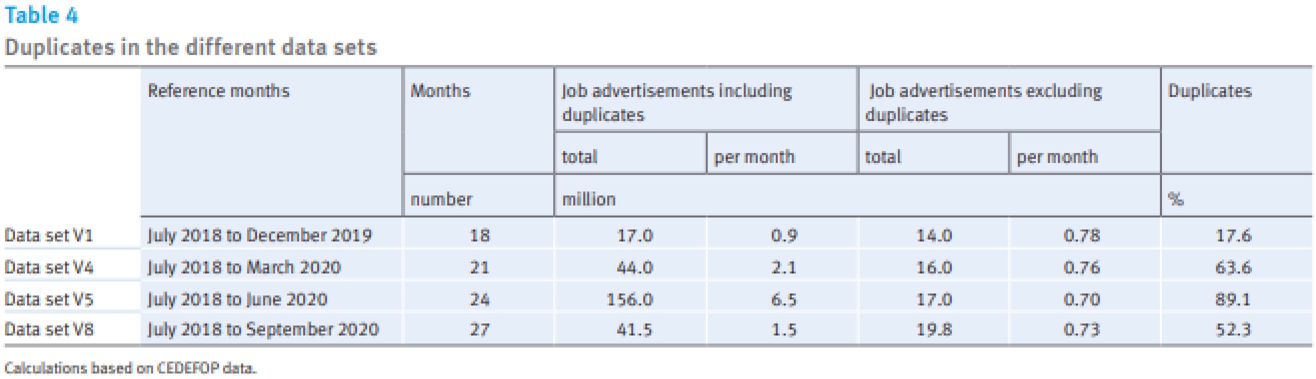


Table 4

Duplicates in the different data sets

	Reference months	Months	Job advertisements including duplicates		Job advertisements excluding duplicates		Duplicates
			total	per month	total	per month	
		number	million				%
Data set V1	July 2018 to December 2019	18	17.0	0.9	14.0	0.78	17.6
Data set V4	July 2018 to March 2020	21	44.0	2.1	16.0	0.76	63.6
Data set V5	July 2018 to June 2020	24	156.0	6.5	17.0	0.70	89.1
Data set V8	July 2018 to September 2020	27	41.5	1.5	19.8	0.73	52.3

Calculations based on CEDEFOP data.

Table copied from: De Lazzer & Rengers (2021), «Impact of the corona virus on the labour market: experimental statistics based on data from online job portals»

Example: Comparability over Time

e.g. WP3, UC1, France about real estate portal «SeLoger.fr»

”The website is loaded with ads posted by realtors, and recently individuals who do not want to use intermediary for their real estate transactions or renting out”

Funded by the European Union

Thank you for your attention!

Magdalena.Six@statistik.gv.at

Alexander.Kowarik@statistik.gv.at

Web Intelligence Network

Funded by the European Union

WIH web scraping: Methodological issues

Piet JH Daas (Statistics Netherlands)
Jacek Maślankowski (Statistics Poland)

23 November 2022

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence
Network

Funded by
the European Union

Outline

- Web scraping:
    - Population perspective
    - Getting URLs
    - Scraping issues (and how to deal with them)

- Concluding remarks

Funded by the European Union

Web scraping:

- Based on an available IT-environment, that can be used for web scraping, a number of methodological challenges emerge
- The majority of them are generic for all scraping processes and need to be dealt with as good as possible (‘learning by doing’)

What are the essential steps in web scraping?	*BREAL function*
1. Identify the target population	*Specify needs*
2. Obtain the URLs of the units where you want to collect data from	‘*Data representation*’
3. Make a request to these URLs to get the associated HTML-page (incl. check robot.txt)	***Acquisition** and recording*
4. (Use specific locators to find the part of interest in the HTML-code)	*Data wrangling*
5. Save the data/page (in some format)	*Acquisition and **recording***

Funded by the European Union

1. Population perspective

- Depending on the topic of interest, various numbers of websites need to be scraped
- Please be aware that not all units of interest may actually have a website!

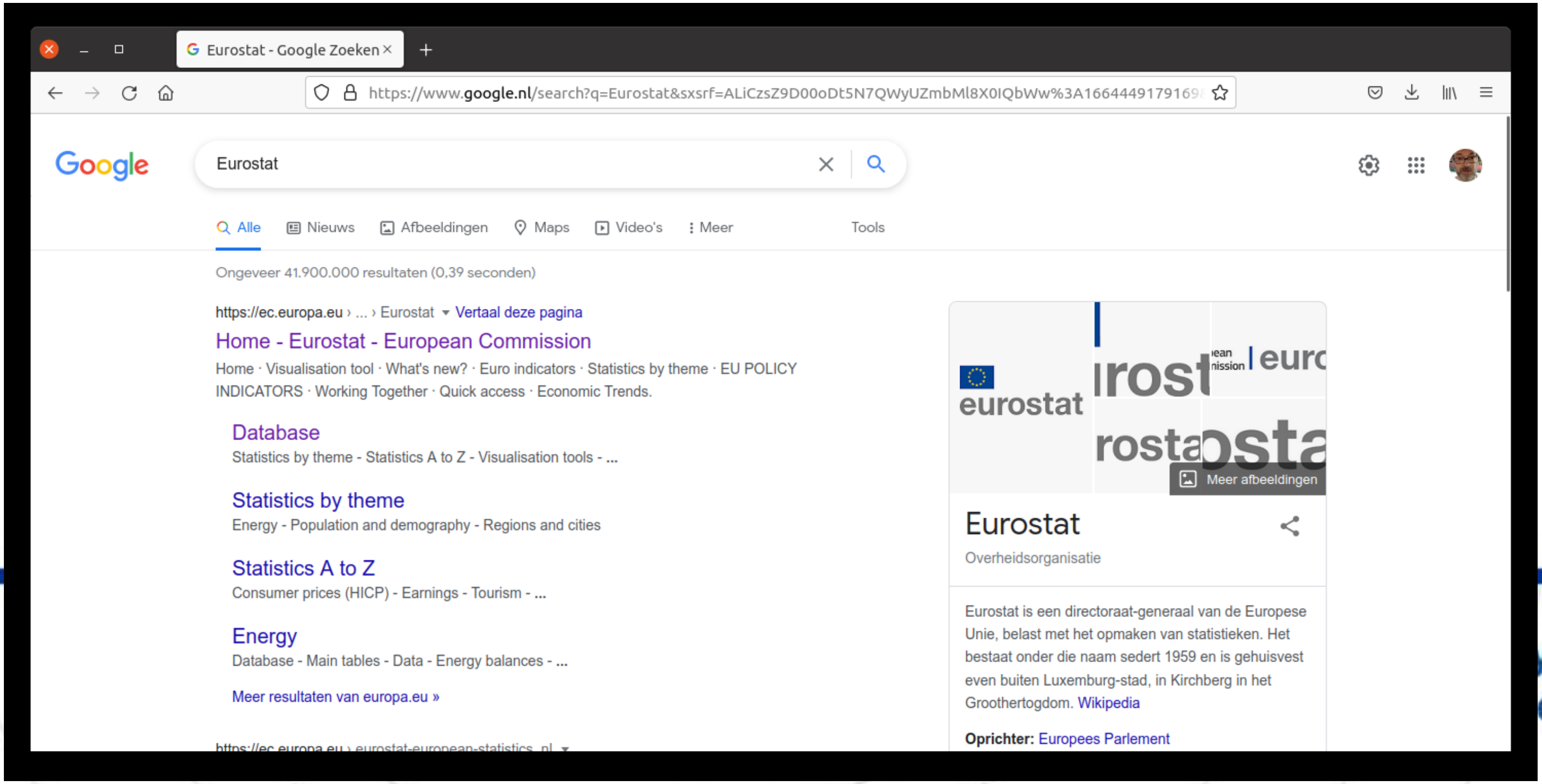
TABLE 1. WEB SCRAPING EXAMPLES BY POPULATION SIZE

Population size	Examples
P1: One website	Satellite data Search engine results
P2: Selected websites (Purposing sampling)	Online Job Advertisements Real estate prices Price statistics
P3: All websites	Enterprise characteristics Innovative company detection

2. Obtaining URLs

- For the units identified, the URL of their website need to identified
  - URL = Uniform Resource Locator (http://www.example.com)

- We can obtain URLs by:
  - Finding and re-using existing lists
  - Buy from others (commercial)
  - Find URLs with URL-finding approach



Eurostat - Google Zoeken
https://www.google.nl/search?q=Eurostat&sxsrf=ALiCzsZ9D00oDt5N7QWyUZmbMl8X0IQbWw%3A166444917916
Google
Eurostat
Alle
Nieuws
Afbeeldingen
Maps
Video's
Meer
Tools

Ongeveer 41.900.000 resultaten (0,39 seconden)

https://ec.europa.eu › ... › Eurostat ▾ Vertaal deze pagina

Home - Eurostat - European Commission

Home · Visualisation tool · What's new? · Euro indicators · Statistics by theme · EU POLICY INDICATORS · Working Together · Quick access · Economic Trends.

Database

Statistics by theme - Statistics A to Z - Visualisation tools - ...

Statistics by theme

Energy - Population and demography - Regions and cities

Statistics A to Z

Consumer prices (HICP) - Earnings - Tourism - ...

Energy

Database - Main tables - Data - Energy balances - ...

Meer resultaten van europa.eu »

https://ec.europa.eu › eurostat-european-statistics_nl ▾

Eurostat

Overheidsorganisatie

Eurostat is een directoraat-generaal van de Europese Unie, belast met het opmaken van statistieken. Het bestaat onder die naam sedert 1959 en is gehuisvest even buiten Luxemburg-stad, in Kirchberg in het Groothertogdom. Wikipedia

Oprichter: Europees Parlement

Funded by the European Union

3. Scrape webpages

- URLs refer to webpages that can be scraped
- However, scraping can be done in various technical ways
  - The technique used may affect if a webpage can be scraped!

- Not all URLs refer to an existing domain/web page, not all URLs may respond, not all URLs refer to a ‘valid’ webpage
- A total of 11 webscraping issues have been identified (on next slides)

TABLE 2. RESULTS OF POLISH CASE STUDY

Specification	Number of websites
Population size	503,700 (100%)
Unique domain names	459,700 (91%)
Accepted connections	340,700 (74%)

No.	Issue	Methods
1	List of URLs is not complete	Use URL search to find additional URLs
2	List of URLs has non-updated data	Use URL search script to verify if URLs have changed
3	Non-recent data on website	Regularly scrape websites
4	Website is blocking robots	Try to use an alternative approach to scrape data and inform website owner of the issue
5	Robots.txt rejection	Inform website owner of intention to scrape the data (scrape anyway)
6	Temporary unavailability	Attempt to scrape website at another time/date
7	No time stamps	Regularly scrape website and monitor changes by comparing stored data in NoSQL database

Funded by the European Union

No.	Issue	Methods
8	Duplicates of websites	Consider de-duplication mechanisms, include URL-forward checks
9	Limited information obtained	Check if website is still active and, if that’s the case, check script to extract more data.
10	The quality of the link between an enterprise and the URL	Check whether the website refers to the enterprise in the population by verifying that company details, like name or address exists in the content of the website.
11	Information on enterprises without a website (if relevant)	Check whether there are other sources of information available, such as a survey, or contact a small sample to obtain an indication of the number of enterprises and type(s) of data missing.

Conclusions, final remarks

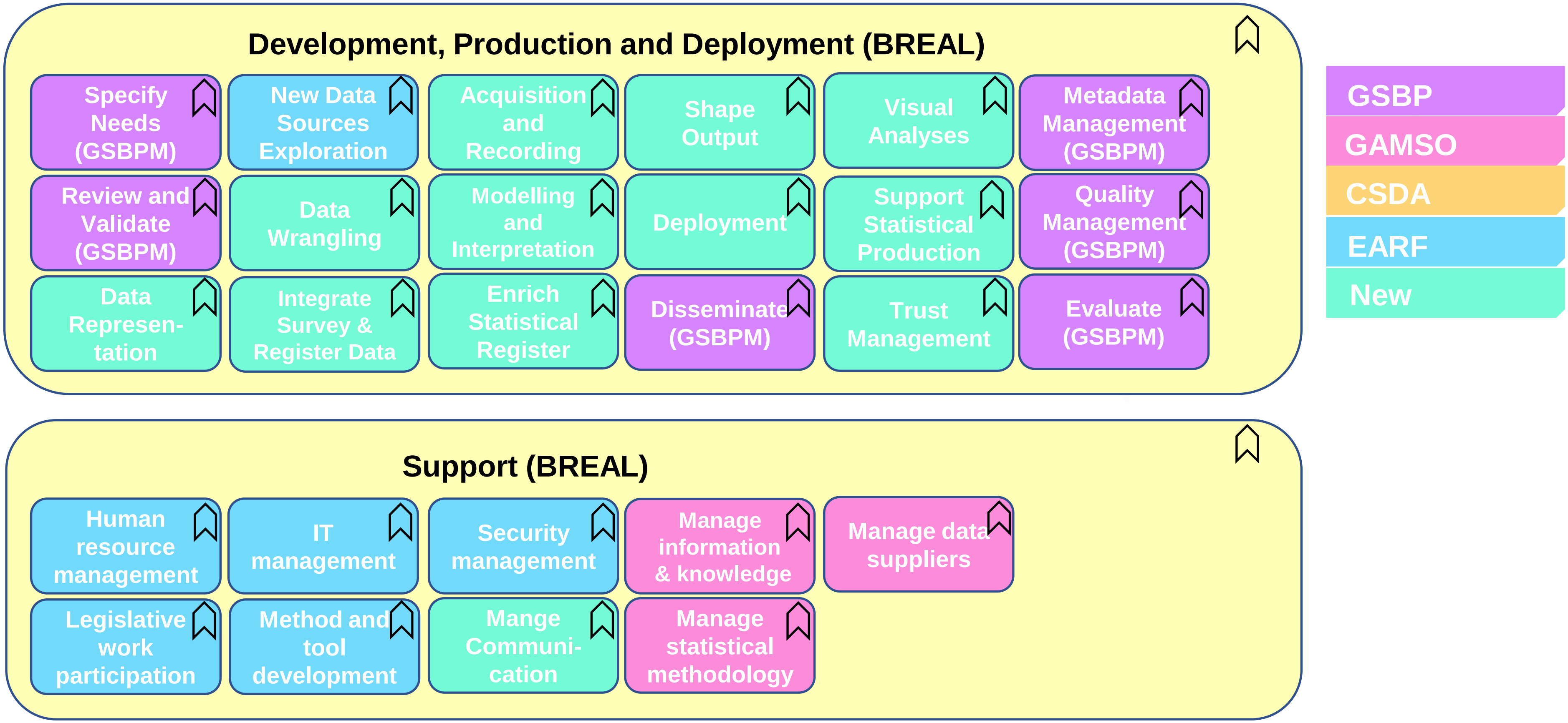
- The web (www) is a very interesting source of data
- The user is in control when collecting data

- There are many applications for which web-data can be used
  - Consumer Price Index (scrape product prices)
  - Real estate info (scrape real estate prices)
  - Vacancy statistics (or similar) (collect online job advertisements)
  - Find websites & URLs (use and collect search engine results)
  - Identifying subpopulation of enterprises (platform economy, innovative comp., ..)
  - Obtain satellite pictures (find and scrape pictures available online)
  - ….

- Methodology is needed to deal with data collection and quality issues as good as possible
  - To produce the best statistics possible
  - General methodology vs. specific methodology ?

Annex: BREAL and WIH implementation

BREAL business functions



Development, Production and Deployment (BREAL)

Specify Needs (GSBPM)	New Data Sources Exploration	Acquisition and Recording	Shape Output	Visual Analyses	Metadata Management (GSBPM)
Review and Validate (GSBPM)	Data Wrangling	Modelling and Interpretation	Deployment	Support Statistical Production	Quality Management (GSBPM)
Data Representation	Integrate Survey & Register Data	Enrich Statistical Register	Disseminate (GSBPM)	Trust Management	Evaluate (GSBPM)

Support (BREAL)

Human resource management	IT management	Security management	Manage information & knowledge	Manage data suppliers
Legislative work participation	Method and tool development	Mange Communication	Manage statistical methodology	

GSBP
GAMSO
CSDA
EARF
New

Source: Scannapieco M., Bogdanovits F., Gallois F.; Fischer B., Kostadin G., Paulussen R., Quaresma S. et al. (2019): BREAL. Big Data Reference Architecture and Layers. Version 2019-12-09. Edited by EUROSTAT

Thank you for your attention!

pjh.daas@cbs.nl