ISI OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

isi

WELCOME.

OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

STATISTIK AUSTRIA
Die Informationsmanager

Statistics and Internet CPS70

Quality Aspects of Web Data

based on the Experiences of the ESSnet Trusted Smart Statistics – Web Intelligence Network

Magdalena Six, PhD,
Magdalena.six@statistik.gv.at
Dr. Alexander Kowarik
Alexander.Kowarik@statistik.gv.at
Statistics Austria

18th of July, 17.30-18.30

Outline/Content

- ESSnet Trusted Smart Statistics
  Web Intelligence Network (WIN)
- "Minimal Guidelines and Recommendations for Implementation" – Structure of document and examples
- Ongoing quality work in the ESSnet:
  - **Landscaping** and selection of websites
  - Quality measures for **hierarchical classifications**

ESSnet Web Intelligence Network (WIN)

ESSnet WIN (2021 – 2025) :

Consortium of 17 organizations from 14 European countries

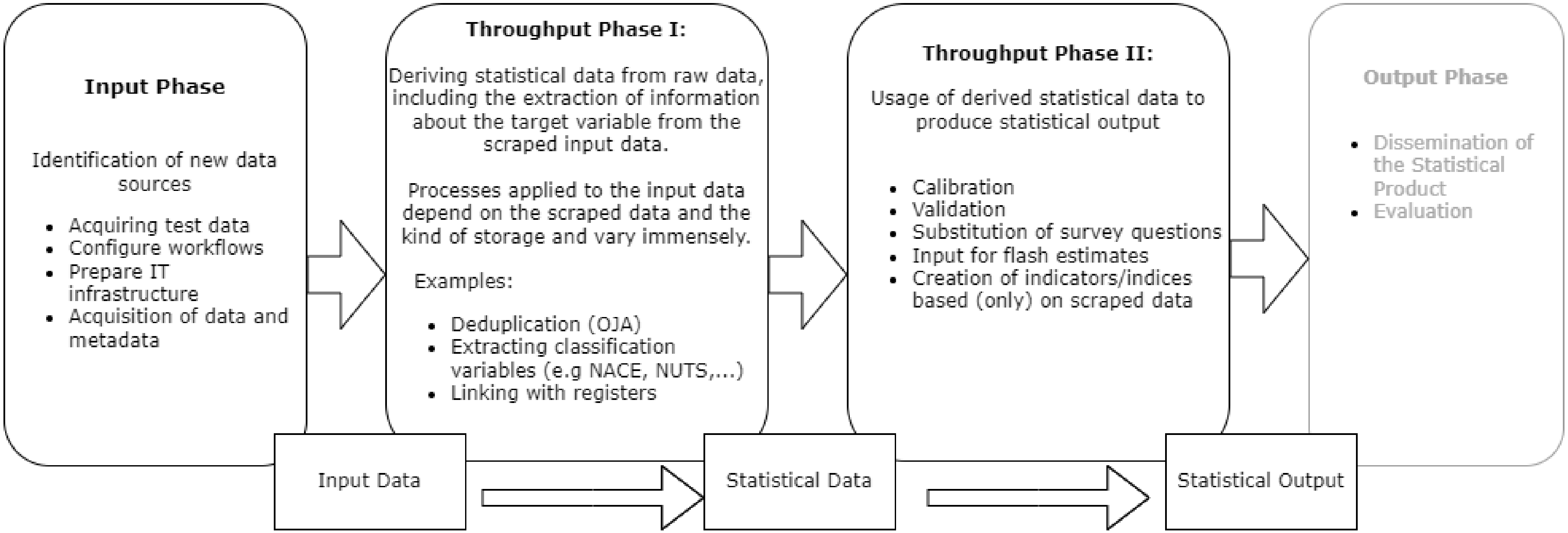
4 Work Packages (WP)

- WP1 Coordination, support an dissemination
- WP2 **Online Job Advertisements** (OJA)
  Online Based Enterprise Characteristics (OBEC)
- WP3 New use cases: real estate, construction activities, prices of household appliances, hotel prices, business register enhancement
- WP4 Methodology and Quality

More information about WIN including a **Blog**:
https://cros-legacy.ec.europa.eu/content/work-package-3-%E2%80%93-new-use-cases_en

„Minimal Guidelines and Recommendations for Implementation“

- Based on existing Quality Guidelines from ESSNet Big Data II
- Structured along the production process:



Input Phase

Identification of new data sources

- Acquiring test data
- Configure workflows
- Prepare IT infrastructure
- Acquisition of data and metadata

Throughput Phase I:

Deriving statistical data from raw data, including the extraction of information about the target variable from the scraped input data.

Processes applied to the input data depend on the scraped data and the kind of storage and vary immensely.

Examples:

- Deduplication (OJA)
- Extracting classification variables (e.g NACE, NUTS,...)
- Linking with registers

Throughput Phase II:

Usage of derived statistical data to produce statistical output

- Calibration
- Validation
- Substitution of survey questions
- Input for flash estimates
- Creation of indicators/indices based (only) on scraped data

Output Phase

- Dissemination of the Statistical Product
- Evaluation

Input Data

Statistical Data

Statistical Output

https://cros-legacy.ec.europa.eu/content/deliverable-41-minimal-guidelines-and-recommendations-implementation_en

„Minimal Guidelines and Recommendations for Implementation“

Guidelines for Throughput Phase I include the following

Quality Aspects:

- Coverage
- Comparability over time
- Measurement errors
- Model errors (from raw data to statistical data)
- Processing errors (from raw data to statistical data)

Examples for Quality Guidelines

Guidelines with respect to Coverage and Representativeness

- Try to estimate the population size and compare with traditional data. For example, when you are scraping enterprise characteristics, try to count the number of websites that are accessible and can be used for web scraping. Compare this number with the data from your business register.

- Make a pilot web scraping to assess what information is included on the websites.
  Check if specific information, e.g. territorial unit or industrial sector, can be extracted from the website. When information on the website is limited, it is also not very likely to monitor enterprise activity (e.g., innovations in enterprises) on the website.

Examples for Quality Guidelines

Guidelines with respect to Comparability over Time

- Check if the modification/update date can be extracted from the website.
- When web-scraping specific information from the website (e.g. job vacancies), try to extract the date of publishing this information.
- When the website is not up to date it is unlikely to detect enterprise activity in longer time series.

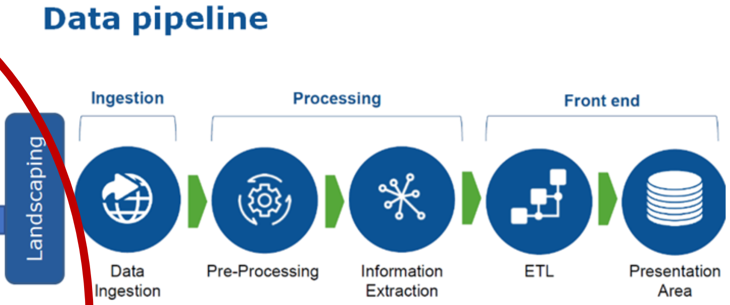
Ongoing quality work of WP4: Landscaping

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 OJA 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)

Figure: Data pipeline and respective quality aspects for OJA, WIH

Ongoing quality work: Landscaping

Definition: **Landscaping** refers to the cataloguing and measurement of all web-based data sources relevant for the topic of interest.

The effort of landscaping varies depending on the topic of interest:

- All needed data might be available on **one website**
  Example: satellite data

- The great extent of existing websites and the impossibility to scrape and combine them all makes it necessary to **select websites**
  Examples: online job advertisements, real estate prices or price statistics

- **All websites** w.r.t. topic of interest should be scraped, combination of ingested information is possible
  Example: enterprise characteristics

Ongoing quality work: Selection of websites

Which websites to scrape?

-> Most important ones? Highest quality?

-> **Score** is needed

Three groups of information to take into account:

- **Information from the website** (stop criteria, mandatory variables, optional variables..)
- **Information about the website** (e.g. market share, rank of Google search, coverage of niche markets, reliable owner of website,...)
- **Experience** (test scraping, prior rounds of scraping)

Ongoing quality work: Selection of websites

Course of action:

- Decide which **groups of information** and which **criteria** to take into account
- Choose a **model** to incorporate all selected criteria to calculate a **score**
- Calculate score and **rank all respective websites**
- Scrape the best-ranked websites
- Document each step and re-evaluate after some time

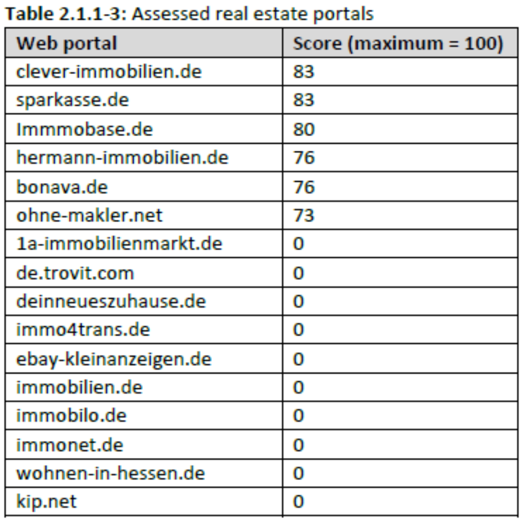


Table 2.1.1-3: Assessed real estate portals

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, UC1, Score for assessed real-estate portals for Germany

Examples

- Members of WP3 „New use cases“ agreed on a score based only on information from website
  https://cros-legacy.ec.europa.eu/content/wp3-deliverable-31-wp3-1st-interim-technical-report-20220330_en

- Eurostat‘s score for ranking OJA websites inlcuded also metainformation and expertise from country experts

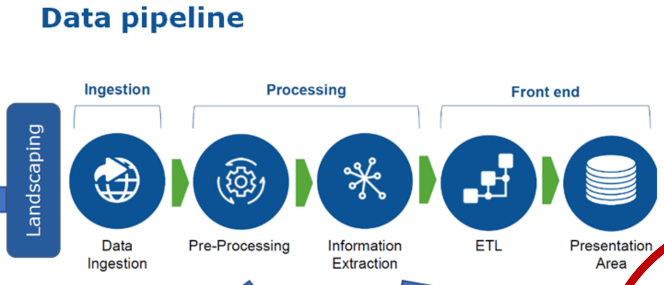
Ongoing quality work: Hierarchical classifications

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 OJA 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)

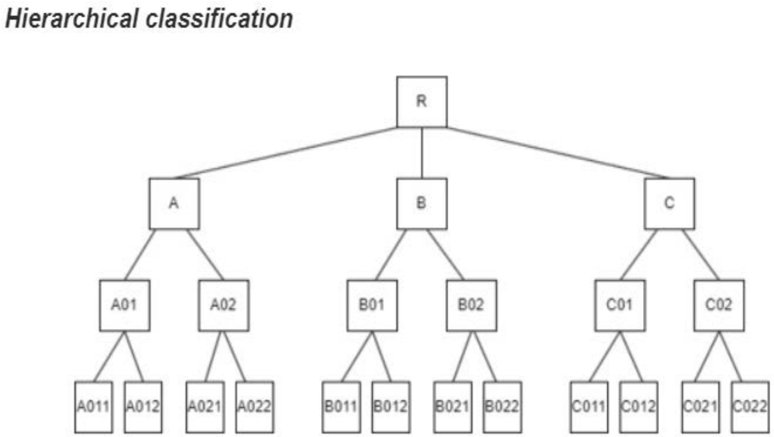
Accuracy for hierarchical classifications

Extracting **hierarchical classification variables** from scraped information is an essential process step. Typical classifications: NACE, NUTS, ISCO …

Traditionally, **accuracy** is measured for **flat classifications**: predicted value is true or false

-> Need for an evaluation measure, which takes into account how **close the predicted and the true value are!**

Overview of **recommended evaluation measures**: https://cros-legacy.ec.europa.eu/content/issue-13-data-accuracy-hierarchical-classification_en



OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

THANK YOU.