Web Intelligence Network Conference
From Web to Data

4-5 February 2025
GDANSK - POLAND

Quality Guidelines for acquiring and using web scraped data

ESSnet WIN, WP4
Magdalena Six, Alexander Kowarik, Manveer Mangat, Johannes Gussenbauer (AUT)

Web Intelligence Network

Funded by the European Union

Outline

- Organisational background
- Statistical production process incl. web-data
- Theoretical Framework for Landscaping
- Examples of quality guidelines in the throughput phase
- Guidelines for a centralized webscraping platform

Funded by the European Union

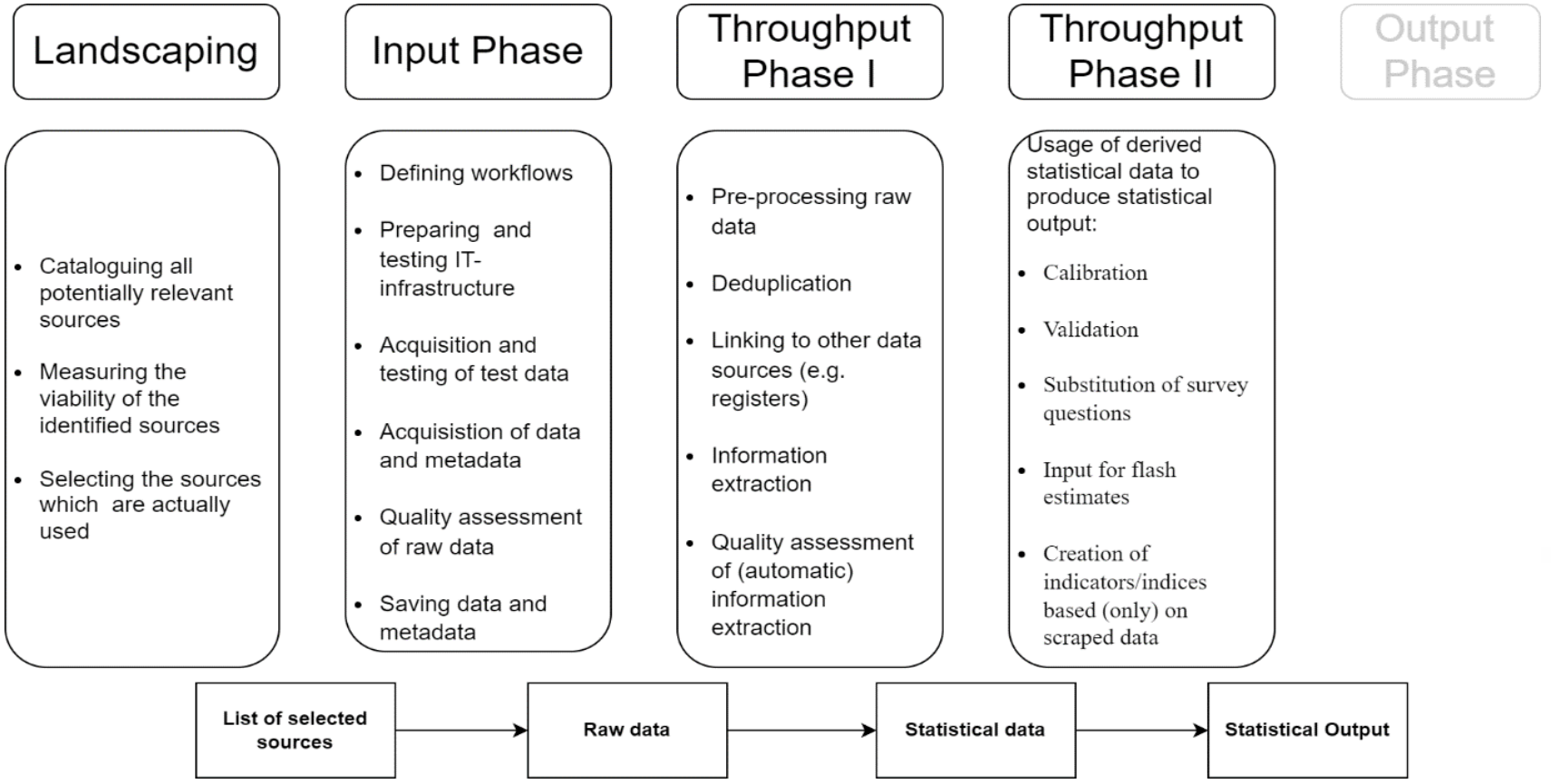
Organisational background

Subgroups of WP4 of ESSnet WIN

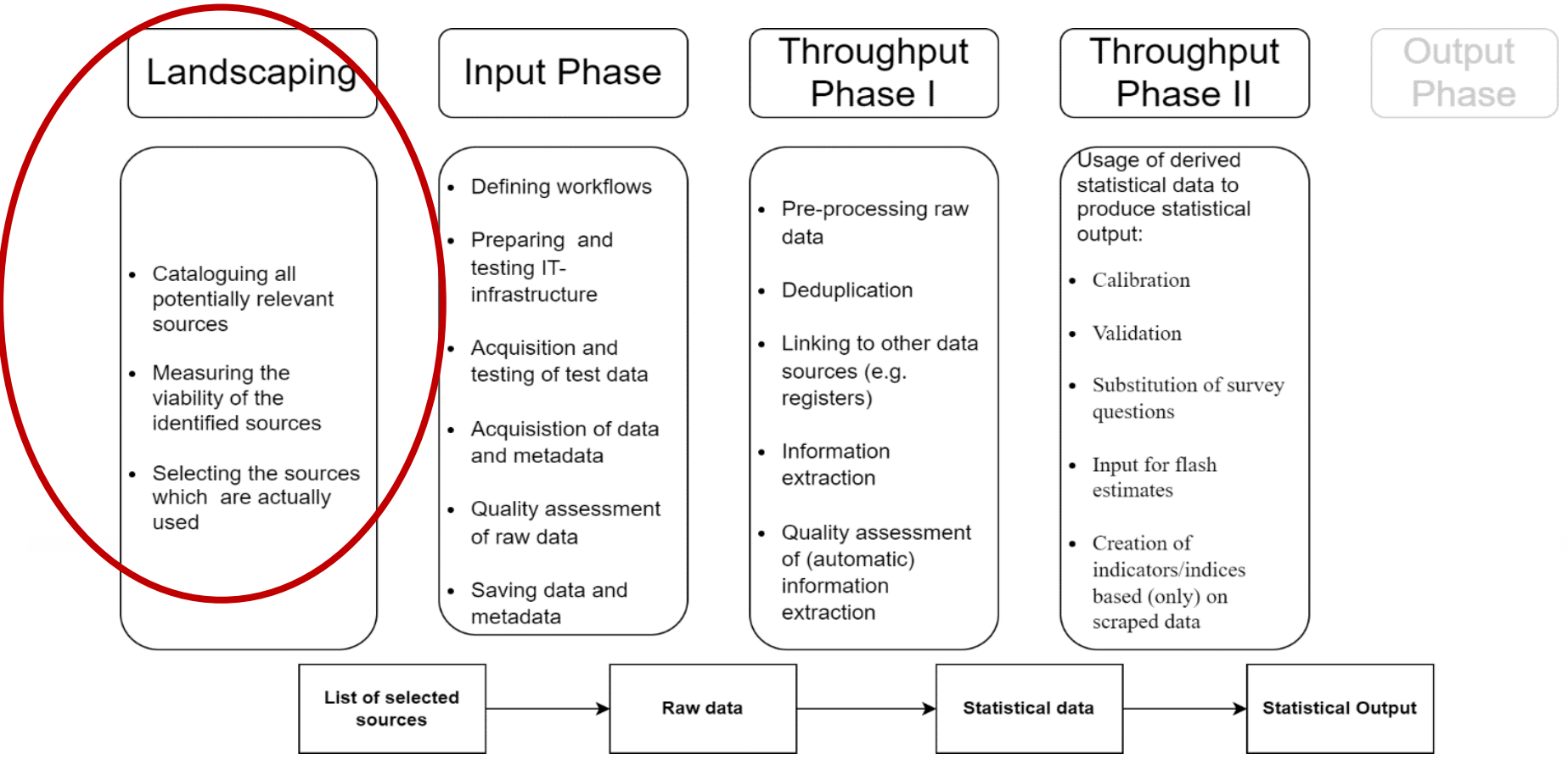
- Methodology
  Deliverable 4.6: WP4 Methodology report on using webscraped data
- Architecture
  Deliverable D4.7: BREAL - Big Data REference Architecture and Layers for web scraped data
- **Quality**
  Deliverable 4.5: Quality Guidelines for acquiring and using web scraped data
- Quality Assessment
  Deliverable 4.8: Quality Assessment for the Statistical Use of Web Scraped Data

All deliverables of WP4 at
https://github.com/WebIntelligenceNetwork/Deliverables

Quality-relevant processes along the production process



Quality-relevant processes along the production process



Spotlight: 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

Landscaping: 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, reliability of owner of website,...)
- **Experience** (test scraping, prior rounds of scraping)

Selection of websites, course of action

Course of action:

- Decide which **groups of information** and which **criteria** to take into account
- Choose a multicriteria decision making **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

Examples

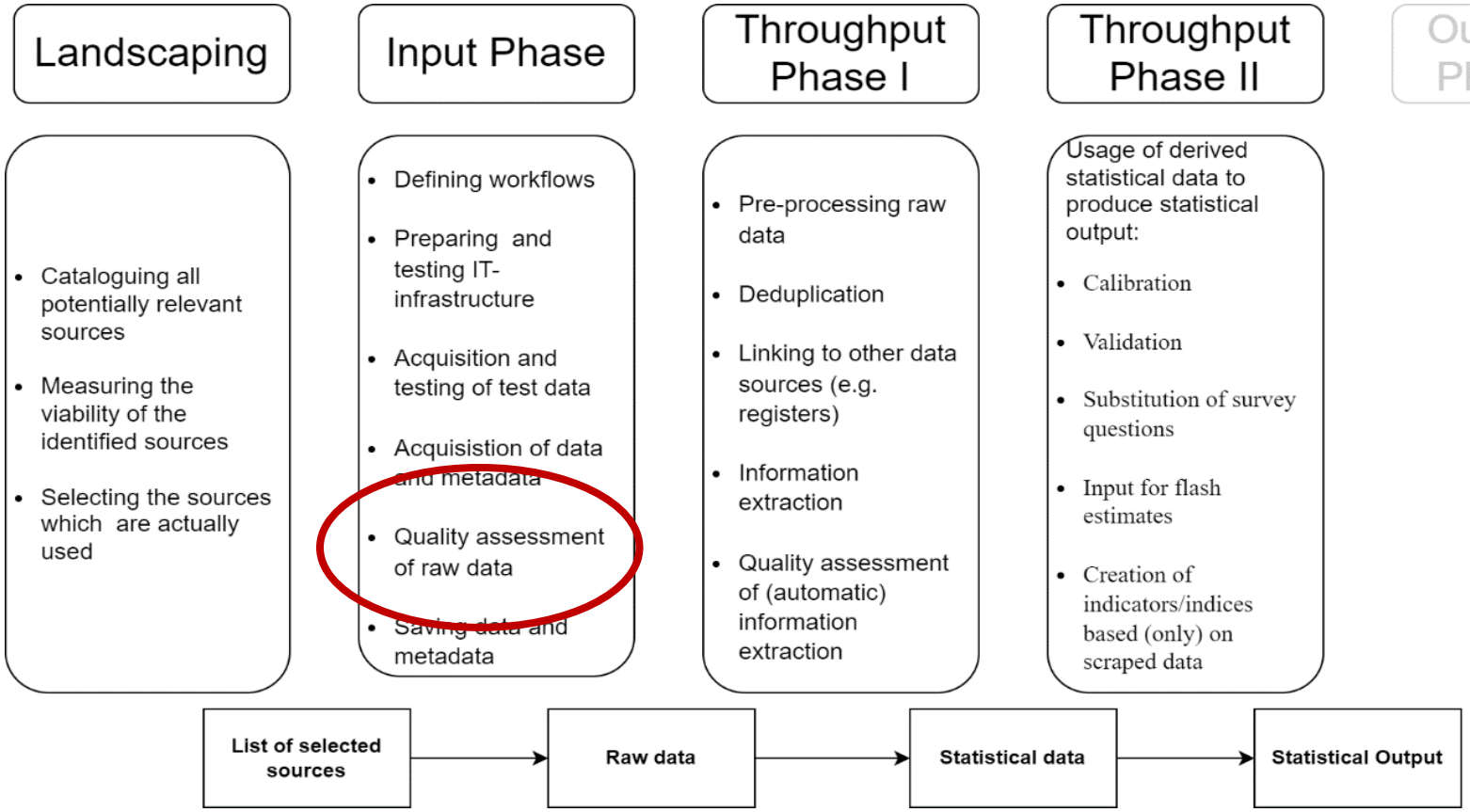
- Members of WP3 „New use cases“ agreed on a score based only on information from website
- Eurostat‘s score for ranking OJA websites inlcuded also metainformation and expertise from country experts

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

Quality-relevant processes along the production process



Spotlight: Example of guidelines for raw OJA data

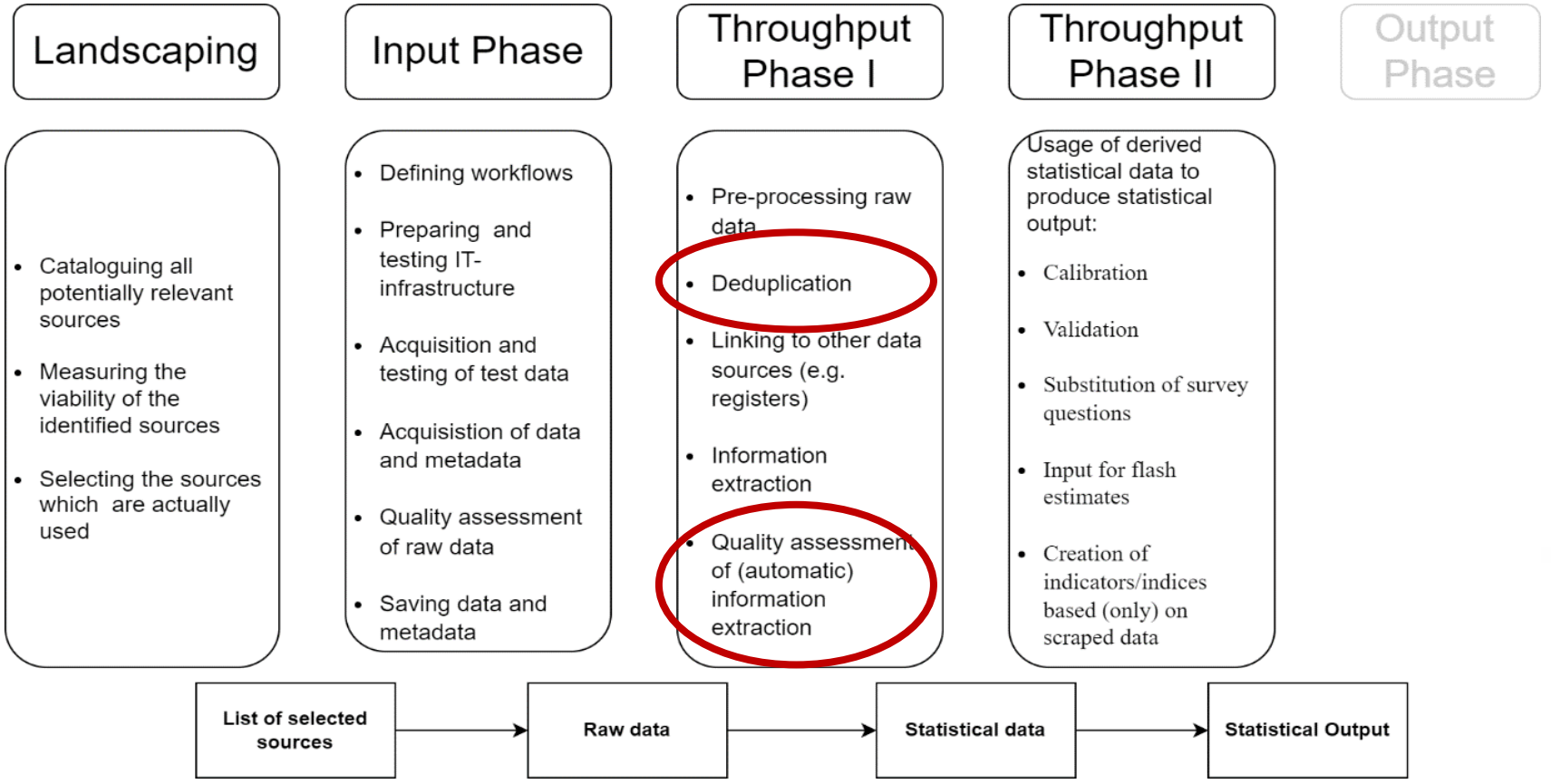
Background: How to detect **concept drift** (representativity over time)?

Even with a constant number of scraped OJA sources, you do not know if an increase (decrease) in the number of scraped OJAs indicates a change in the job market or a change in the popularity of the source.

Proposed guidelines to measure changes in the popularity of the sources:

- Calculate the ranking of the most important sources w.r.t the OJA volume and observe this ranking over the course of time
- Determine the number of OJAs per source and check (e.g. via a plot of the individual time series) if the dynamics of the individual time series per source are similar

Quality-relevant processes along the production process



Spotlight: Example of guidelines about deduplication

Background: Duplicates of scraped entities (offers, advertisements,..) lead to overcoverage.

Proposed guidelines for duplicates within one source and across sources

- Describe your deduplication strategies for duplicates within a source
  Example WP3 UC2 Germany: Deduplication within a portal uses the portal's own ID for an ad or object: only the last / newest ad is kept. Same objects can have different Ids, thus as second step we check address and other characteristics

- Describe your deduplication strategies for duplicates across sources
  Example WP3 UC2 Germany: treat all offers at same address as duplicates gives lower bound, deduplication needs heuristic approaches because despite the use of satellite data, it is not deterministically possible to decide always if two offers are duplicates

Spotlight: Guidelines for annotation exercises

Background: Annotation is the process of manually labelling or classifying data to validate results from an automatic information extraction process. Given the volume of the data the annotation is usually done on a selected sub sample of the whole available data.

Guidelines for assessing the quality of a specific classification by annotating a sample

- *Design the sample according to the needs*
- *Determine the necessary sample design*
- *Define time horizon*
- *Establish annotation guidelines*

Bonus: Guidelines for a centralized scraping infrastructure

Guidelines about technical requirements of a centralized web data infrastructure

- *Smooth operation of the scraping processes*
- *Portability*
- *Open-Source-First*
- *Modularity*
- *Web-native, user-friendly access modality*
- *Metadata for a transparent, traceable scraping process*
- *Scheduling, Prioritising and Resource management*
- *...*

Thank you for your attention!

Funded by the European Union