

# Navigating quality challenges in landscaping web data: new aspects and source stability

Magdalena Six and Alexander Kowarik  
Statistics Austria, Center for Methodology

Rome, 7 December 2023

[www.statistik.at](http://www.statistik.at)

Independent statistics for evidence-based decision making

# Quality aspects for setting up a scraping process

## Webbased information – just scrape it!?

**(Jr.) Statistiker, Mathematiker, Data Scientist (m/w/d)** 

**Unternehmen:** smartPM.solutions GmbH  
**Arbeitsort:** 1070, Wien,Neubau  
**Arbeitszeit:** Teilzeit/Vollzeit  
**Inseriert/Aktualisiert:** 20.09.2023  
**Dienstverhältnis:** ArbeiterInnen/Angestellte  
**Berufsgruppe:** MathematikerIn, Data Scientist (m/w)

**Kurzbeschreibung:**  
in Teilzeit 25h Und/ Oder Vollzeit DEINE AUFGABEN:  
Als Teil unseres Projektteams sind Sie im Bereich Business-Analytics beschäftigt und setzen sich mit Prognosen, Analysemodellen und Simulationen auseinander. Nach genauen Anforderungsanalysen arbei...

**Technischer Assistent:in**

**Unternehmen:** Medizinische Universität Wien  
**Arbeitsort:** 1090, Wien,Alsergrund  
**Arbeitszeit:** Teilzeit  
**Inseriert/Aktualisiert:** 13.09.2023  
**Dienstverhältnis:** ArbeiterInnen/Angestellte  
**Berufsgruppe:** Projektassistentin in der Forschung, Projekttechnikerin

**Kurzbeschreibung:**  
An der Medizinischen Uni für Medical Data Science Statistik mit der Kennzahl eines/einer halbeschäftigt Assistent:in (gemäß Kollektive Verwendungsgrup...

**Praktikum im Bereich Audit**

**Unternehmen:** Oesterr Nationalbank  
**Arbeitsort:** 1090, Wien,Alsergrund  
**Arbeitszeit:** Teilzeit/Vollzeit  
**Inseriert/Aktualisiert:** 25.09.2023  
**Dienstverhältnis:** ArbeiterInnen/Angestellte  
**Berufsgruppe:** Aushilfskraft (m/w)

**Kurzbeschreibung:**  
Dauer von 6 Monaten (Bezahlung 100%, ab 01. Jänner 2024 Vereinbarung möglich Reue Ihre Aufgaben: + Unterstützen und quantitativen Analyse...

**SAMSUNG CU8070 (2023) 50 Zoll Crystal UHD Smart TV**

**Produkttyp:** LCD TV  
**Bildqualität:** UHD 4K  
**Bildschirmdiagonale (cm/Zoll):** 125 cm / 50 Zoll  
**Betriebssystem:** Tizen™ Smart TV

★★★★★  
 Vergleichen

**PHILIPS 32PHS6808/12 (2023) 32 Zoll HD-ready Smart TV**


**Produkttyp:** LCD TV  
**Bildqualität:** HD-ready  
**Bildschirmdiagonale (cm/Zoll):** 89 cm / 32 Zoll  
**Betriebssystem:** Smart TV mit neuem Betriebssystem

UVP 329,-  
**€ 211,-**  
inkl. MwSt. versandkostenfrei

● Online verfügbar  
Lieferung Mi, 10.10.2023 - 05.10.2023

● Abholung  
Bitte wählen Sie einen Markt aus: **Markt auswählen**

★★★★★  
 Vergleichen

**Booking.com** EUR  ? Ihre Unterkunft anmelden

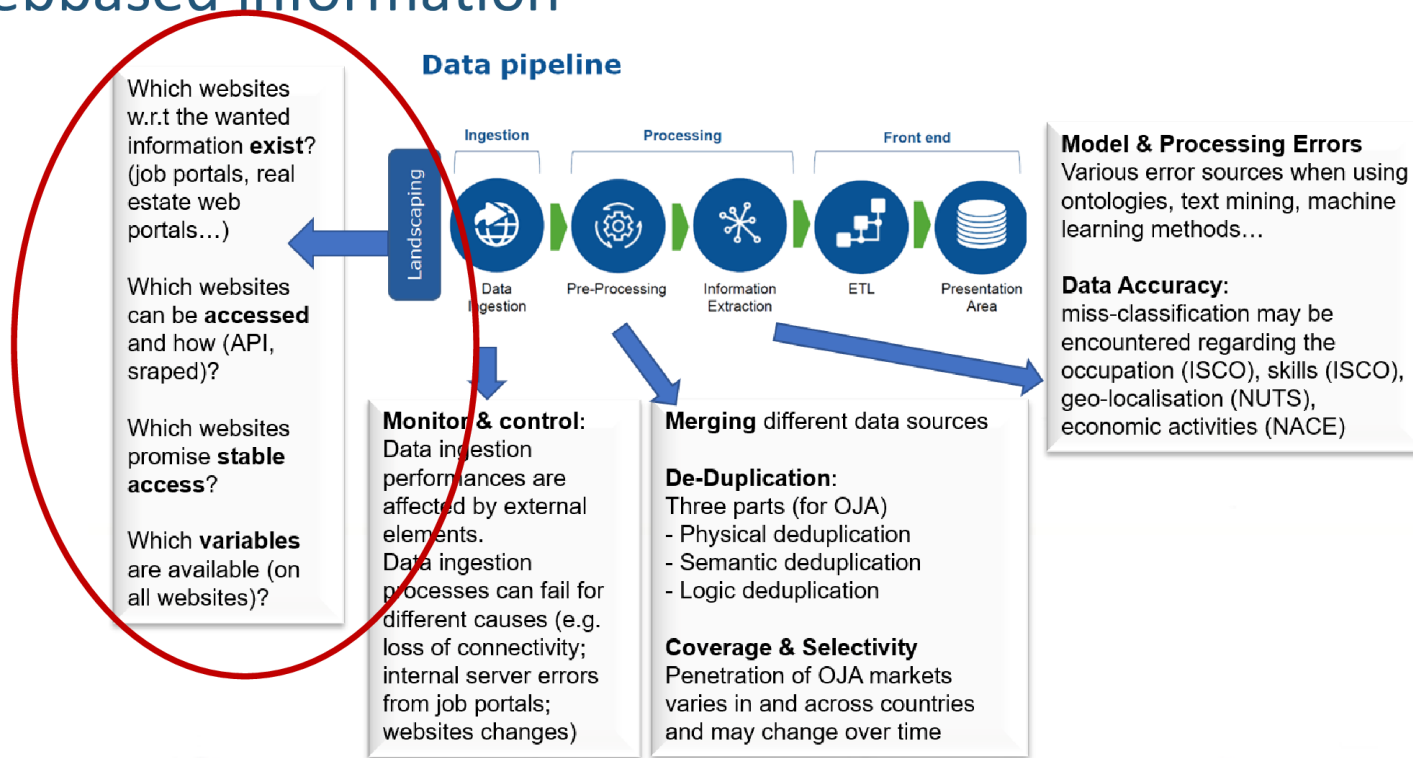
## Finden Sie Ihre nächste Unterkunft

Finden Sie Angebote für Hotels, Ferienunterkünfte und vieles mehr ...

×

Ich suche ganze Ferienunterkünfte oder -wohnungen  Ich reise geschäftlich  Ich suche auch nach Flügen

# Example of data pipeline (OJA) and quality aspects for webbased information

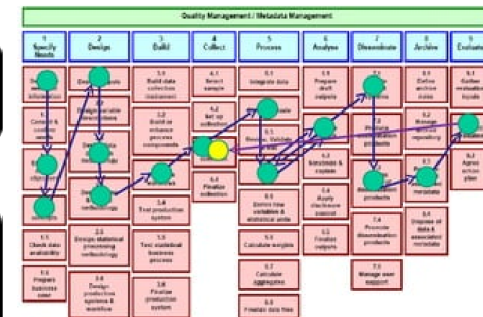
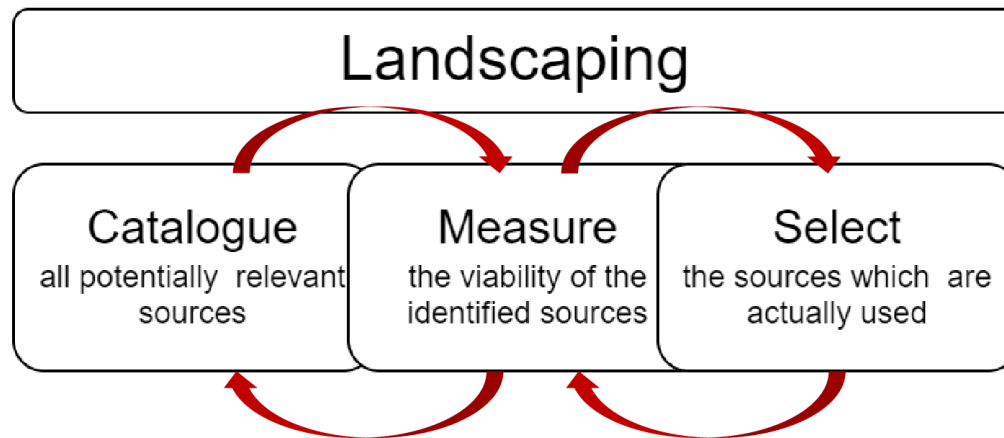


Data pipeline from ESSnet WIN, WP2, Use Case Online Job Advertisements

# Landscaping

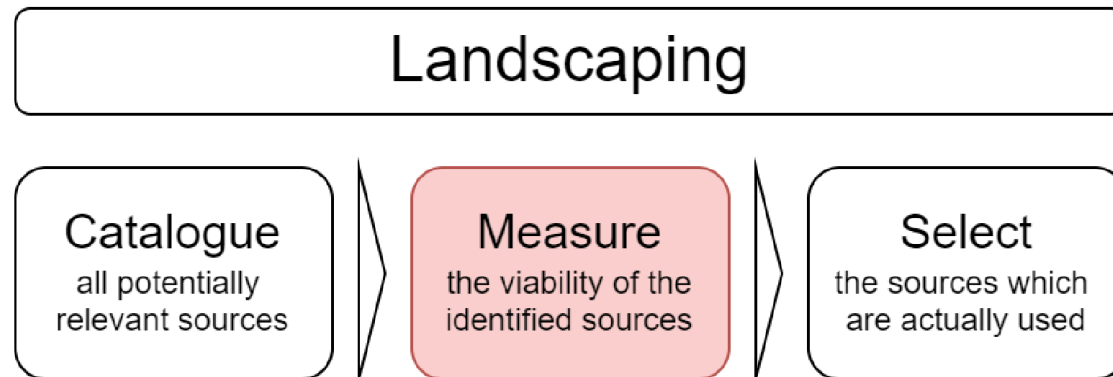
## Definition (Attempt)

**Definition:** **Landscaping** comprises all process steps necessary to **catalogue** all relevant sources for a specific topic of interest, to **measure** the quality and technical viability of the catalogued sources and to **select** the sources, which are actually used, based on the measured criteria.



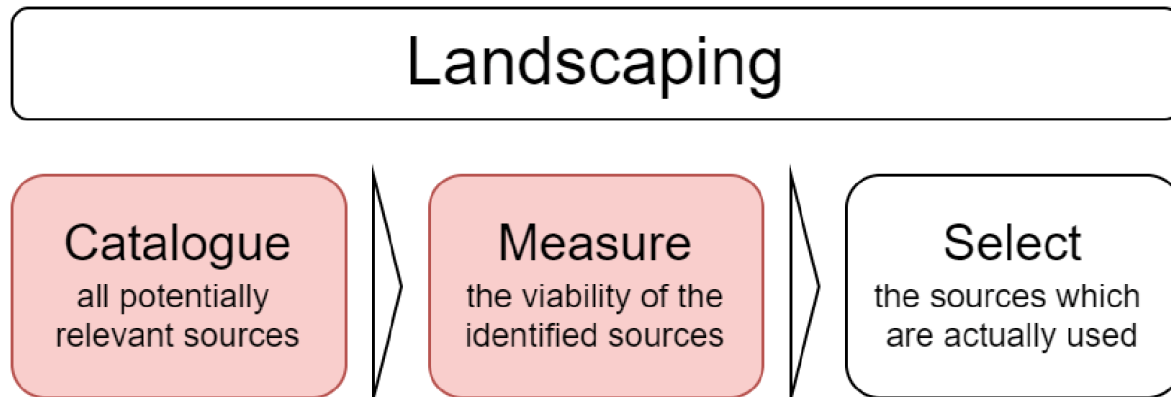
# Varying complexity for varying topic of interests

Examples – Scrape prices from a price comparison website (e.g. „Tarifkalkulator“)



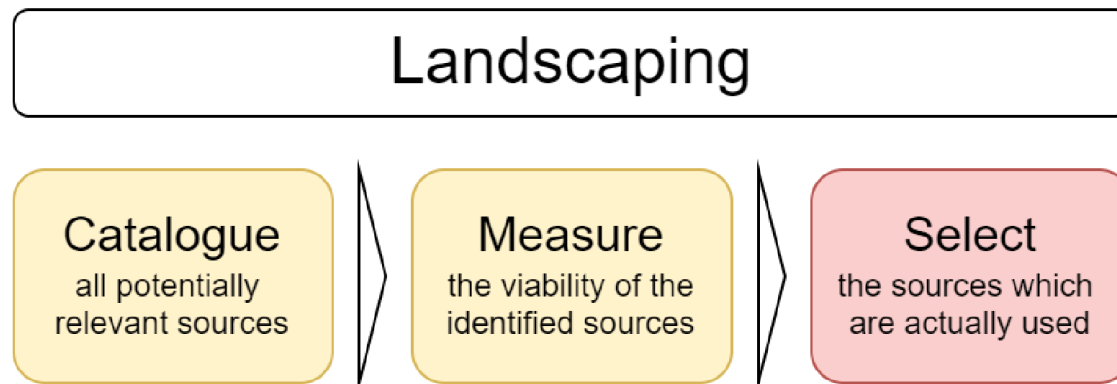
## Varying complexity for varying topics of interest

Examples: „Web-based information about the European drone industry”,  
“Online-Based Enterprise Characteristics”



# Varying complexity for varying topic of interests

## Example: Online Job Advertisements (OJA)



# Catalogue

The cataloguing process very much depends on the following questions:

- Does your topic of interest require to **find all websites** or only a **list of representatives** which fulfil certain criteria?
- Do you have any **“additional” information**, for example a list of companies whose websites you want to scrape?

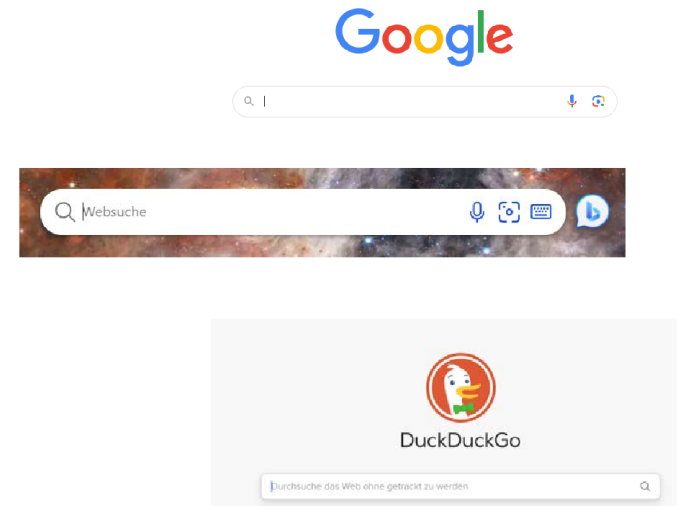
	<b>Find all websites</b>	<b>Find representatives</b>
<b>Additional information, apart from websites, available</b>	Online Based Enterprise Charactersitics (existene of homepage, online sale, social media usage) <i>Additional information: Business Register</i>	Prices of clothes on online-shopping websites  <i>Additional Information: Companies with hightes turnover from business register</i>
<b>Only information from the respective websites available</b>	Identify the population of businesses active in a sector (e.g. green industry, drone industry)	Online Job Advertisements (centralised scraping by Eurostat)



# Catalogue

**Attention**, the search results **depend** very much on:

- the exact **terminology** searched for
- **the search engine** used (Google, Bing, DuckDuckGo,..)
- the location (**IP address**) of the searching user,
- the user's previous search history (**cookies**),
- the **country extension** of the search engine used (e.g. using Google.nl vs. Google.ie)
- the User-agent of the 'browser' program used.



Daas et al (2022) list the following ways **to reduce these effects**:

- using a **VPN connection**,
- using a browser that has no search history or searching the web via an anonymised (**incognito**) browser, and
- using a search engine that is specific to the country under study.

Piet Daas et al. Web intelligence for measuring emerging economic trends: the drone industry, Statistical Working Papers, 2022

# Measure

Broad subprocess, varies w.r.t. topic of interest, question about all websites/representatives...

- Learning about technical aspects of website (e.g. Captcha, robots blocking)
- Scraping test runs
- Classification models w.r.t to question: Does catalogued website belong indeed to topic of interest? (e.g. Is this indeed the homepage of a company active in the drone industry?)
- Identifying information on catalogued websites which can be matched to business register e.g. VatNr („UID“ , „FN“)
- Collect information about criteria needed for selection models (overlap with „Select“-subprocess)
- ....

# Handover to Magdalena



# Selection of Websites

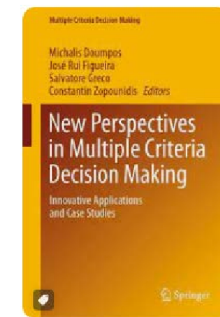
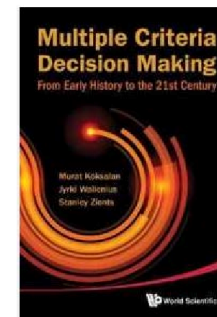
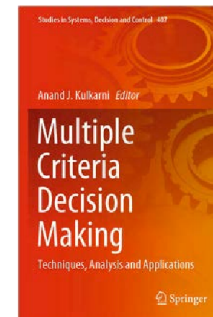
Which websites to scrape (from the list of catalogued websites)?

- Most important ones? Highest quality?
- Example for Online Job Advertisements (OJA) sources:

	Popularity	Reliability of Owner/Operator	Number of needed variables available in structured fields	Original Ads?
Jobportal 1	High	High	0	Yes
Jobportal 2	Medium	Low	4	Yes
Jobportal 3	High	Low	2	No

-> Multicriteria Decision Making Problem!

- Alternatives to be ranked
- Criteria
- Model which produces ranking of alternatives based on values of alternatives for each criteria



# MCDM – Which Criteria? Which Method (Model)?

**Three groups of criteria** to take into account:

- Information **from the website** (technical criteria, mandatory variables, optional variables, structured vs not-structured information, up-to-date information, volume of information..)
- Information **about the website** (e.g. market share, rank of Google search, coverage of niche markets, information from business registers, reliability of owner of website,...)
- **Experience** (test scraping, prior rounds of scraping, stability of results)

## **Models ?**

Extended weighted sum model

Analytical hierarchy process (AHP) model

# Selection of Websites

## Course of action:

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

# Example for Selection Models

ESSnet WIN, WP3 New Use Cases: Scraping prices of real estate market, hotel prices, online prices of household appliances

## Criteria (only from website)

- **Stop criteria and minimal criteria** for all use cases:  
Robots blocking , Captcha, filter criteria, up-to-date content,..
- **Mandatory variables** per use case: e.g. ad\_id, building type, number of rooms, price, square meter, brand of product, energy efficiency class, city, adress....
- **Optional variables** per use case: e.g. rent furnished, buidling year, parking ,elevator, distance in km to city center, Ratings based on user reviews, parking for clients...)

## Selection Model:

Score = 0 if at least one stop criteria, minimal criteria or mandatory variable not fulfilled

Score = Sum over all fulfilled criteria, and scaled to 100 (max)

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

# Example for Selection Model: ESSnet WP2, Online Job Advertisements

## Final score constructed from two building blocks

1. a quantitative assessment of adherence of each website to the desired characteristics  
**Model: Analytical Hierarchical Process (AHP)**
2. qualitative assessment of the sources' relevance (ranking by country-experts, model?)

## Criteria included in first building block (AHP score)

- the type of the job-portal (primary job portal, secondary job portal or mixed),
- the type of the operator (classified ads portal, company websites, national newspaper, recruitment agency, ...),
- the OJA volume displayed on the website,
- the sectoral scope (one or more),
- the displayed form (structural field, text or mixed) for variables such as “Type of Occupation”, Type of Contract”, “Working Time” etc,
- ...



# Example for Selection Model: ESSnet WP2, Online Job Advertisements

Criteria included in the second building block (International Country Experts Rank)

- **Popularity** was measured by the websites' relative interest as produced by Google Trends
- **Stability** involved several criteria, affecting the **stability of the access to the website** as well as the **stability of the time series** based on the scraped data.
- **Coverage** refers to the question if the scraped OJAs **cover all groups belonging to a classification of interest** such as ISCO or NUTS in a **similar way as comparable known data**.

Combining the two building blocks - the AHP score and the ICEs' ranks – leads then to a **final score**



## Example: centrally scraped OJA data, Quality aspects for „provided“ data

Looking at the sources from a user's  
perspective

# Relevance and stability of source stability

- Indicators for relevance of selected sources
  - Name the 5 most important job portals in your country.
  - Check if the most important job portals are included in the list of relevant sources
  - Check if the most important job portals are included in all of the years
- Indicators for the stability of existence of the included sources
  - Is it the same sources in the data over time?
  - Check if most important sources are included at several points in time.
- Indicators for the stability of the popularity of the included sources
  - Ranking w.r.t. volume of OJAs per source / stability of ranking of sources
  - Check development per source
- Indicators for the stability of sources over different versions of data
- *Quality of the estimated classification -> “**annotation exercises**”*



## Summary

- This is part of the deliverable **‘Quality Guidelines for acquiring and using web scraped data’**.
- A draft version can be shared informally.
- Everything is still under development.
  
- Check the source stability in existing data sets
- Systematically think about which sources should be included