STATISTICS AUSTRIA
The Information Manager

Johannes Gussenbauer
Alexander Kowarik
Lukas Mikesa
Marlene Weinauer
Jakob Peterbauer
Quality Management and Methodolgy (QM)

ESSNet WPC

Webscraped data for replacing and validating survey questions

UNECE Statistical Data Editing
September 2020

Overview

- Main Concepts and Data
- Setup
- Results
  - Website
  - Social Media presence
  - Webshop

Main Concepts and Data

- ESSNET Big Data II WPC
  - Online Based Enterprice Characteristics (OBEC)
  - Documentation and Results
  - Source Code on github

Main Concepts and Data

- ESSNET Big Data II WPC
  - Online Based Enterprice Characteristics (OBEC)
  - Documentation and Results
  - Source Code on github

- Use data scraped from webpages (or collected through APIs) to get information (ICT Usage Survey) on units from a statistical buisness register
  - Does Enterprise have a webpage?
  - Uses Social Media?
  - Has a webshop?
  - ...

Setup

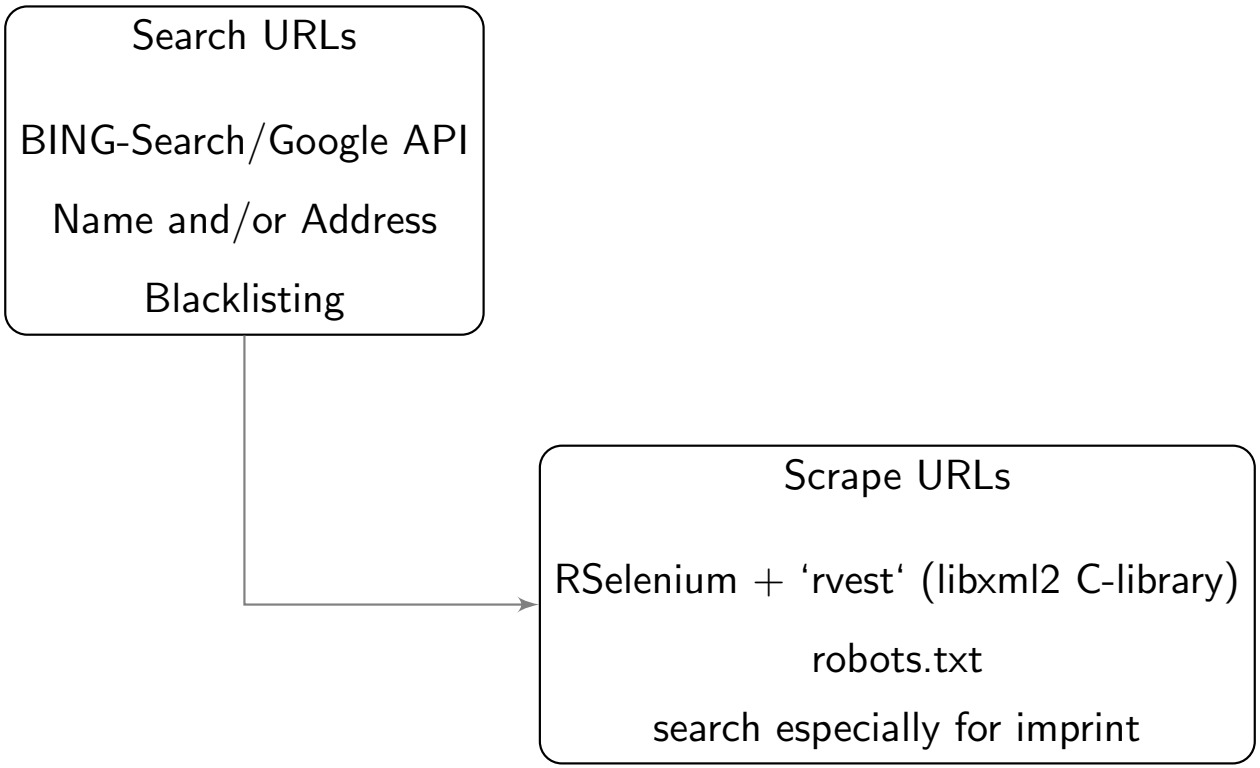
Search URLs

BING-Search/Google API

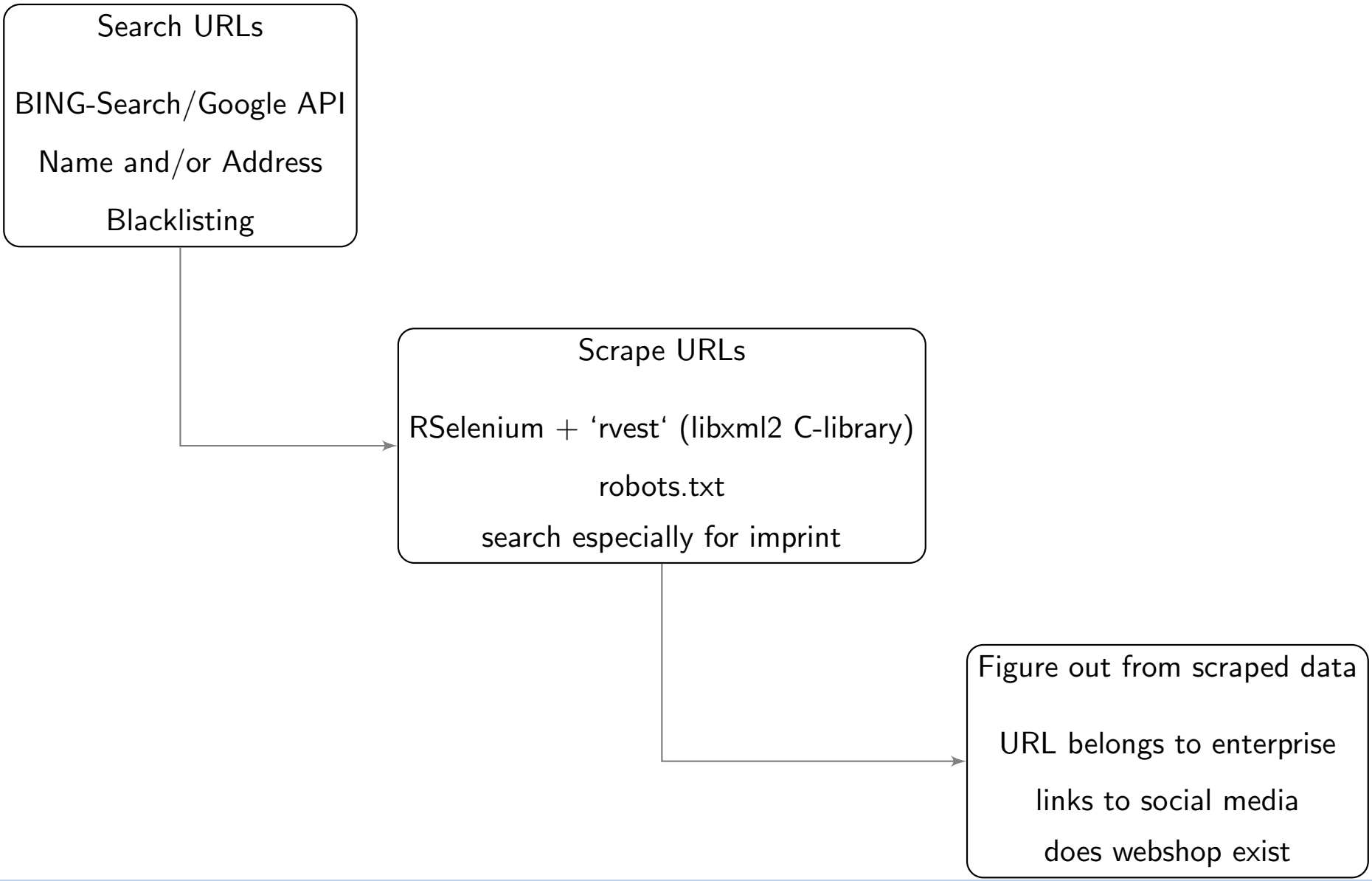
Name and/or Address

Blacklisting

Setup


Search URLs
BING-Search/Google API
Name and/or Address
Blacklisting
Scrape URLs
RSelenium + ‘rvest‘ (libxml2 C-library)
robots.txt
search especially for imprint

Setup


Search URLs
BING-Search/Google API
Name and/or Address
Blacklisting
Scrape URLs
RSelenium + ‘rvest‘ (libxml2 C-library)
robots.txt
search especially for imprint
Figure out from scraped data
URL belongs to enterprise
links to social media
does webshop exist

Remarks on Results when comparing with ICT Survey

- Comparing with ICT $\rightarrow$ respect enterprise groups

This observation variable doesn't refer specifically to the ownership of the website, but to the use of a website by the enterprise to present its 'business'. It includes not only the existence of a website which is located on servers belonging to the enterprise or are located at one of the enterprise's sites, but also third party websites (e.g. one of the group of enterprises to which it belongs i.e. website of the parent company or holding company).

Remarks on Results when comparing with ICT Survey

- Comparing with ICT $\rightarrow$ respect enterprise groups

This observation variable doesn't refer specifically to the ownership of the website, but to the use of a website by the enterprise to present its 'business'. It includes not only the existence of a website which is located on servers belonging to the enterprise or are located at one of the enterprise's sites, but also third party websites (e.g. one of the group of enterprises to which it belongs i.e. website of the parent company or holding company).

- Comparison uses unweighted results

URL Validation

URL Validation

- Try to identify enterprise in scrapped text

- Currently two ways to link enterprise with URL
  1. VAT and or company register number are found
  2. Name and/or Address are found
     - build simple model from information found
     - predict 0/1 variable

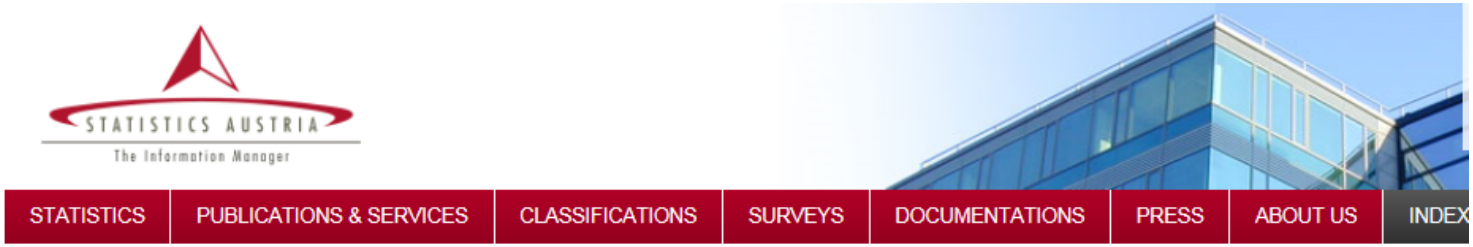
URL Validation

- Try to identify enterprise in scrapped text

- Currently two ways to link enterprise with URL
  1. VAT and or company register number are found
  2. Name and/or Address are found
     - build simple model from information found
     - predict 0/1 variable

- Austrian Media Act §25
  - Enterprise should state VAT and or company register number (CRN) as well as Name, Adress,... on website
  - Link to this information on main page and any sublink

URL Validation



Website information

Media owner

STATISTICS AUSTRIA
Federal Institution under Public Law
Guglgasse 13
A-1110 Vienna
Tel.: +43 (1) 71128 0
Fax: +43 (1) 71128 7728
office@statistik.gv.at

Company register: FN 191155k, registry court: Vienna Commercial Court
Registered office: Vienna, place of jurisdiction: Vienna
VAT ID No.: ATU37869909

Data Protection Information:

- www.statistik.at

dsgvo@statistik.gv.at

Disclosure in accordance with § 25 Austrian Media Act

URL Validation - Results

URL Validation - Results

Year		URL rate
2019	ICT Survey	93.66
2020	ICT Survey	93.94
2019	Webscraping	82.44
2020	Webscraping	91.18

			Webscraping	
			0	1
ICT	2019	0	123	51
		1	359	2212
	2020	0	83	84
		1	160	2427

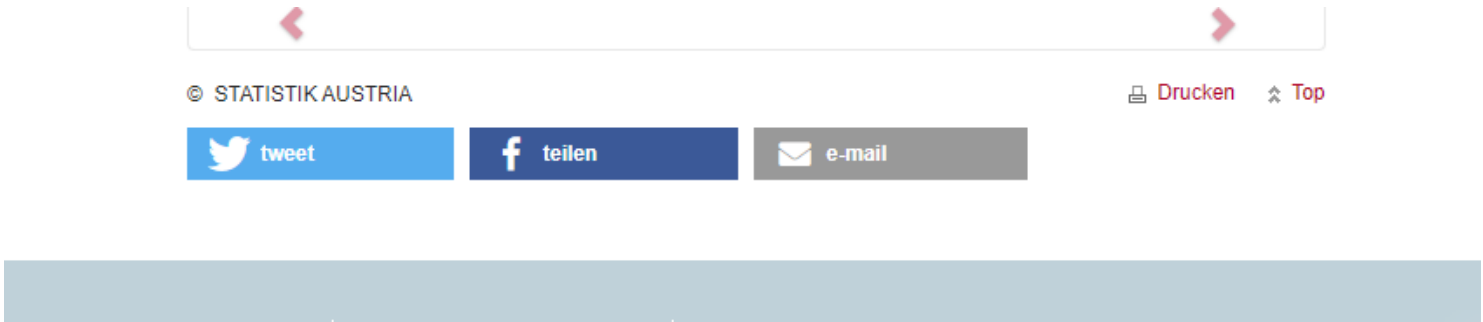
Website: Issues

- Scrapped URL corresponds to parent company abroad(!)
  - No Imprint / foreign VAT (Germany, Switzerland)
- robots.txt forbids scraping imprint
- Not all companies adhere to Austrian Media Act §25

Social Media presence

Social Media presence

- Search for links to social media sites on web page



- Drop links which referre to legal notice, policy, ect. . . (`https://de-de.facebook.com/policies/ads`)

Social Media presence - Resuts

Social Media presence - Resuts

Year		URL rate
2019	ICT Survey	51.22
2020	ICT Survey	57.23
2019	Webscraping	50.16
2020	Webscraping	62.71

			Webscraping	
			0	1
ICT	2019	0	964	375
		1	404	1002
	2020	0	740	438
		1	287	1289

Social Media presence - Issues

- FALSE NEGATIVES:
  - conceptional reasons: what is social media
  - too deeply nested

- FALSE POSITIVES:
  - Reference to social media profile which is not owned by the enterprise

Social Media presence - Issues

- FALSE NEGATIVES:
  - conceptional reasons: what is social media
  - too deeply nested

- FALSE POSITIVES:
  - Reference to social media profile which is not owned by the enterprise

→ How to validate the social media profile?

Web Shop presence

Web Shop presence

- Search for Core words
  - expert-based approach
  - modelling approach (Random Forest, Classification Tree, Naive Bayes)
- For modelling approach

$$has\ webshop \sim core\_word_1 + core\_word_2 + \ldots + core\_word_p$$

Web Shop presence - Results

Year		URL rate
2019	ICT Survey	28.16
2020	ICT Survey	32.64
2019	Webscraping	25.17
2020	Webscraping	32.57

			Webscraping	
			0	1
ICT	2019	0	1692	280
		1	362	411
	2020	0	1506	349
		1	351	548

Web Shop presence - Issues

- overall poor sensitivity rates (expert-based + modelling approach)
  - expert-based-approach: ~60%
  - modelling approach: ~45% and less
- core-words NACE-dependent
  - hotels do not use ''shopping chart" but 'booking' $\rightarrow$ booking in German however is also beech (a tree)
- NACE-dependend model necessary

Conclusion

- Indicators derived from webscraped data not quite good enough
  - although rate of enterprises having a webpage quite close
- Improve logic for social media indicator
- Improve model for webshop
- Use OBEC from scraped data for validation for ICT 2021

Please address queries to:
Johannes Gussenbauer

Contact information:
Guglgasse 13, 1110 Vienna
phone: +43 (1) 71128–7327
Johannes.Gussenbauer@statistik.gv.at

ESSNet WPC

Webscraped data for replacing and validating survey questions