STATISTICS AUSTRIA
The Information Manager

Johannes Gussenbauer
Alexander Kowarik
Lukas Mikesa
Marlene Weinauer
Jakob Peterbauer
Qualitätsmanagement und Methodik (QM)

Sofia
June 2019

ESSNet WPC

URL-Validation & Social Media presence

Overview

- Setup
- URL Searcher
- URL Validation
- Social Media presence

Setup

- Webscrapping Server
  - outside of STAT Firewall

- Search for URLs with
  - URL-Searcher from ISTAT / BING-Search
  - (Yellow Pages API)

- Webscrapping and Textparsing with R

URL-Searcher

- First 10 Hits from ISTAT URL-Searcher
- Blacklist URLs
  - yellow pages, facebook, business catalogues, ...
  - URL occurs in more than 10 different BING searches
  - currently 123 URLs excluded
  - Not easily extendable, valid URLs can get blacklisted
- Exclude blacklisted URLs from BING-Search Results
- For the ICT
  - Sample 5200
  - Results from BING-Search considering blacklisted URLs 4690

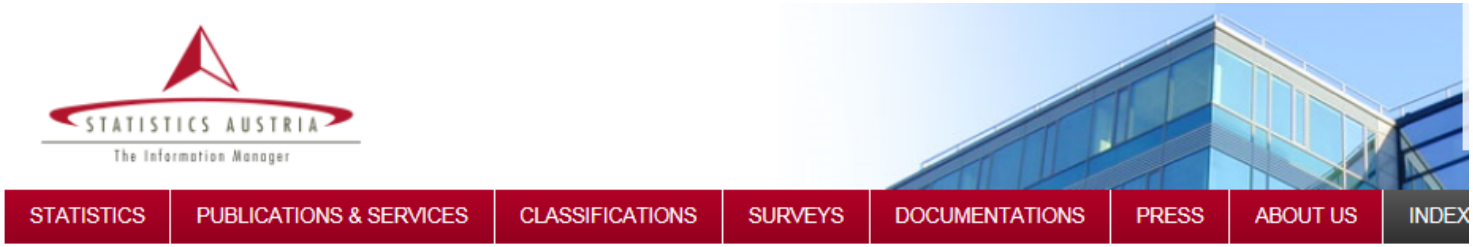
URL Scrapping

- Scrape URL using Selenium (R-package RSelenium)
- Step by step procedure
  1. Respect robots.txt (for now?)
  2. Go to URL
  3. Get all links on homepage having the same base URL
  4. Search especially for links containing “contact”, “legal information”,. . .
  5. Scrape main page (raw HTML) and sublinks (especially contact information)
     - maximum of 25 sublinks + contact-links
  6. Save results in seperate Files - no DB yet
- Repeat steps 1-6 for each enterprise and URL

URL-Validation

- Load scrapping results and parse them using R-Package `rvest` (libxml2 C-library)
- Try to identify enterprise in scrapped text
- Austrian Media Act §25
  - Enterprise must state VAT and or company register number (CRN) as well as Name, Adress,... on URL
  - Link to this information on main page and any sublink
- URL is defined valid if at least one of the 2 informations is found
  - VAT and or CRN
  - full name and adress as listed in our register

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

Number of valid URLs

URL-Status	Number Enterprises	Share in %
found & valid	3215	61.46052
found & not valid	1483	28.35022
not found	533	10.18926

URL valid by VAT/CRN and or Name, Adress,...

		Name/Adress		
		TRUE	FALSE	
VAT/CRN	FALSE	371	1483	1854
	TRUE	1267	1577	2844
		1638	3060	4698

URL-Validation

Comparing with ICT (preliminary results)

		Survey		
		TRUE	FALSE	
Webscrapping	TRUE	1761	17	1778
	FALSE	574	96	670
	NA	163	72	235
		2498	185	2683

URL-Validation Issues

- Scrapped URL corresponds to parent company
  - Search URL for XYC Logistics $\rightarrow$ get URL for XYC AG
  - No direct information on XYC Logistics found at URL

From Eurostat Manual

> *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).*

- Not all companies adhere to Austrian Media Act §25

Social Media presence

- Search for links to social media sites on web page
- Flag if link is found and save link
- Drop links which referre to legal notice, policy, ect. . . (https://de-de.facebook.com/policies/ads)

		Survey		
		1	0	
Webscrapping	TRUE	820	208	1028
	FALSE	244	585	829
		1064	793	1857

Social Media presence - Issues

- Quality of response variable questionable
  - link to social media platform on homepage $\leftrightarrow$ survey response indicates no social media usage

- Reference to social media profile which is not owned by the enterprise
  - How to check the social media profile?

Please address queries to:
Johannes Gussenbauer

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

ESSNet WPC

URL-Validation & Social Media presence