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 Enterprise Characteristics (OBEC)
  - Documentation and Results
  - Source Code on github
Main Concepts and Data

- ESSNET Big Data II WPC
  - Online Based Enterprise 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 business 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 → 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 → 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

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, Address, .. on website

Link to this information on main page and any sublink
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, Address,... on website
  - Link to this information on main page and any sublink
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.: ATU37669509

Data Protection Information:
› www.statistik.at
» dsgvo@statistik.gv.at

Disclosure in accordance with § 25 Austrian Media Act
<table>
<thead>
<tr>
<th>Year</th>
<th>URL Validation Method</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>ICT Survey</td>
<td>93.66</td>
</tr>
<tr>
<td>2020</td>
<td>ICT Survey</td>
<td>93.94</td>
</tr>
<tr>
<td>2019</td>
<td>Webscraping</td>
<td>82.44</td>
</tr>
<tr>
<td>2020</td>
<td>Webscraping</td>
<td>91.18</td>
</tr>
</tbody>
</table>
## URL Validation - Results

<table>
<thead>
<tr>
<th>Year</th>
<th>Method</th>
<th>URL rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>ICT Survey</td>
<td>93.66</td>
</tr>
<tr>
<td>2020</td>
<td>ICT Survey</td>
<td>93.94</td>
</tr>
<tr>
<td>2019</td>
<td>Webscraping</td>
<td>82.44</td>
</tr>
<tr>
<td>2020</td>
<td>Webscraping</td>
<td>91.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Webscraping</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>0</td>
</tr>
<tr>
<td>2019</td>
<td>0</td>
</tr>
<tr>
<td>2019</td>
<td>1</td>
</tr>
<tr>
<td>ICT</td>
<td>0</td>
</tr>
<tr>
<td>2020</td>
<td>1</td>
</tr>
</tbody>
</table>
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
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)
<table>
<thead>
<tr>
<th>Year</th>
<th>URL</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019 ICT Survey</td>
<td>51.22</td>
<td></td>
</tr>
<tr>
<td>2020 ICT Survey</td>
<td>57.23</td>
<td></td>
</tr>
<tr>
<td>2019 Webscraping</td>
<td>50.16</td>
<td></td>
</tr>
<tr>
<td>2020 Webscraping</td>
<td>62.71</td>
<td></td>
</tr>
</tbody>
</table>
### Social Media presence - Results

<table>
<thead>
<tr>
<th>Year</th>
<th>Method</th>
<th>URL rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>ICT Survey</td>
<td>51.22</td>
</tr>
<tr>
<td>2020</td>
<td>ICT Survey</td>
<td>57.23</td>
</tr>
<tr>
<td>2019</td>
<td>Webscraping</td>
<td>50.16</td>
</tr>
<tr>
<td>2020</td>
<td>Webscraping</td>
<td>62.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Web</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>ICT</td>
<td>964</td>
<td>375</td>
</tr>
<tr>
<td>2019</td>
<td>Web</td>
<td>404</td>
<td>1002</td>
</tr>
<tr>
<td>2020</td>
<td>ICT</td>
<td>740</td>
<td>438</td>
</tr>
<tr>
<td>2020</td>
<td>Web</td>
<td>287</td>
<td>1289</td>
</tr>
</tbody>
</table>
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
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
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

<table>
<thead>
<tr>
<th>Year</th>
<th>URL rate</th>
<th>Year</th>
<th>URL rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>ICT Survey</td>
<td>2019</td>
<td>Webscraping</td>
</tr>
<tr>
<td></td>
<td>28.16</td>
<td></td>
<td>25.17</td>
</tr>
<tr>
<td>2020</td>
<td>ICT Survey</td>
<td>2020</td>
<td>Webscraping</td>
</tr>
<tr>
<td></td>
<td>32.64</td>
<td></td>
<td>32.57</td>
</tr>
</tbody>
</table>

## Webscraping

<table>
<thead>
<tr>
<th>Year</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>1692</td>
<td>280</td>
</tr>
<tr>
<td>2020</td>
<td>1506</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>351</td>
<td>548</td>
</tr>
</tbody>
</table>
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’ → 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
ESSNet WPC
Webscraped data for replacing and validating survey questions