Kinne, Jan; Axenbeck, Janna

Working Paper

Web mining of firm websites: A framework for web scraping and a pilot study for Germany

ZEW Discussion Papers, No. 18-033

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Kinne, Jan; Axenbeck, Janna (2018) : Web mining of firm websites: A framework for web scraping and a pilot study for Germany, ZEW Discussion Papers, No. 18-033, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

This Version is available at:
https://hdl.handle.net/10419/181864

Discussion Paper No. 18-033

Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany

Jan Kinne and Janna Axenbeck

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 18-033

Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany

Jan Kinne and Janna Axenbeck

Web Mining of Firm Websites:
A Framework for Web Scraping and a Pilot Study for Germany

Jan Kinne[1,2]* and Janna Axenbeck[3]

[1] Department of Economics of Innovation and Industrial Dynamics, ZEW - Centre for European Economic Research, L7 1, 68161 Mannheim, Germany

[2] Department of Geoinformatics – Z_GIS, University of Salzburg, 5020 Salzburg, Austria

[3] Department of Digital Economy, ZEW - Centre for European Economic Research, L7 1, 68161 Mannheim, Germany

* Correspondence: jan.kinne@zew.de; Phone: +49 621 1235-297

August 2018

Abstract: Nowadays, almost all (relevant) firms have their own websites which they use to publish information about their products and services. Using the example of innovation in firms, we outline a framework for extracting information from firm websites using web scraping and data mining. For this purpose, we present an easy and free-to-use web scraping tool for large-scale data retrieval from firm websites. We apply this tool in a large-scale pilot study to provide information on the data source (i.e. the population of firm websites in Germany), which has as yet not been studied rigorously in terms of its qualitative and quantitative properties. We find, inter alia, that the use of websites and websites' characteristics (number of subpages and hyperlinks, text volume, language used) differs according to firm size, age, location, and sector. Web-based studies also have to contend with distinct outliers and the fact that low broadband availability appears to prevent firms from operating a website. Finally, we propose two approaches based on neural network language models and social network analysis to derive firm-level information from the extracted web data.

Keywords: Web Mining; Web Scraping; R&D; R&I; STI; Innovation; Indicators; Text Mining.

JEL Classification: O30, C81, C88

Acknowledgments: The authors would like to thank the *German Federal Ministry of Education and Research* for providing funding for the research project (TOBI - Text Data Based Output Indicators as Base of a New Innovation Metric) of which this study is a part. Special thanks are due to Georg Licht who contributed valuable help and advice.

Author Contributions: Janna Axenbeck and Jan Kinne designed the study. Jan Kinne gathered, pre-processed, analyzed and visualized the data. Janna Axenbeck and Jan Kinne wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

1. Introduction

The World Wide Web (Web) is a ubiquitous medium for communicating and disseminating information. Billions of private and commercial users worldwide (OECD 2017) are producing increasing amounts of data. However, the sheer amount of data available, along with its mostly unstructured nature and its decentralized storage, impose specific requirements on the collection, pre-processing, and analysis of the data. *Web mining*, the application of data mining techniques to uncover relevant data characteristics and relationships (e.g. data patterns, trends, correlations) from unstructured web data, has been shown to be applicable in many fields of research (Raymond and Blockeel 2000; Askitas and Zimmermann 2015).

In economic research, firm websites are a particularly interesting area of the Web. Firms use their websites to present themselves, as well as their products and services. The information found on these websites can be used to assess firms' products, services, credibility, achievements, key personnel decisions, strategies and relationships with other firms (Gök, Waterworth, and Shapira 2015). Surveying firms using their websites instead of conducting interviews or questionnaires or using other traditional methods, offers clear advantages (scale, cost, timeliness of the survey), but also comes with its own challenges (data collection and harmonization, data analysis). As yet, no consistent approach for studying firm websites has been established. In addition, the data source itself (i.e. the population of firm websites) has not been studied rigorously in terms of its qualitative and quantitative properties. Basic yet important data characteristics such as the structural properties of firm websites and their coverage of the overall firm population are unknown. Using the specific example of innovation in firms, we outline a general framework for extracting information from firm websites using web scraping and data mining which can be adapted to a variety of economic research questions. We also present *ARGUS* (an *Automated Robot for Generic Universal Scraping*), an easy and free-to-use web scraping tool which allows for large-scale information retrieval from websites without requiring the user to have expert knowledge of web scraping technology. The tool is applied in a case study with the aim of deriving best practice guidelines for the large-scale web scraping of firm websites and addressing the following two research questions:

RS1. *URL Coverage*: What subpopulation of firms can be surveyed using web mining of firm websites? Is a systematic bias in terms of firm characteristics (age, size, sector, location etc.) to be expected?

RS2. *Website Characteristics*: How do firm websites differ in terms of their size and content?

The remainder of this paper is organized as follows. First, after identifying the shortcomings of traditional innovation indicators, we outline our approach in developing a web-based innovation indicator (Section 2). In Section 3, we summarize previous findings regarding web-based innovation indicators. In Section 4, we present our data. Section 5 describes the applied ARGUS web scraper. In section 6, we present the results of a large-scale case study and key facts on the website properties of firms based in Germany, which will then be discussed in further detail in Section 6. Section 7 concludes and outlines future research.

2. Web-based Innovation Indicators

Innovation is defined as the implementation of a new or significantly improved product or process and is considered as a main driver of economic growth. The disruptive force of radical innovation has the ability to reshape the economy and pave the way for new periods of long-term economic growth, while incremental innovation causes continuous change. It is therefore a matter of public interest to measure innovation activities within a STI (science, technology and innovation) system. Measuring these innovation activities to a sufficient degree of accuracy allows us to analyze a system's driving factors as well as the effectiveness of STI policies. However, there is evidence that traditional indicators of innovation (e.g. questionnaire-based surveys and patent-based indicators) struggle to provide a timely and sufficiently granular picture of the current state of STI systems (Nagaoka, Motohashi, and Goto 2010; Squicciarini and Criscuolo 2013; OECD 2009). As an example, insufficient innovation indicators are assumed to be one possible cause for the so-called productivity paradox of an accelerating STI system and a simultaneous productivity slowdown as observed in developed countries (OECD 2015).

In this chapter, we identify four major shortcomings of established innovation indicators (questionnaire-based, patent-based, and literature-based) and present a general analysis framework for *web-based innovation indicators*, which have the potential to overcome these shortcomings. Such indicators can be a useful addition to the range of indicators available. We argue that innovation indicators generated from the web mining of firm websites have the potential to offer timely and highly granular information on the STI system and are obtainable on a large scale and at low costs.

2.1. Shortcomings of established inovation indicators

Firm-level innovation is often measured by means of indicators constructed using data from large-scale questionnaire-based surveys. Examples of such surveys include the Oslo Manual-based biennial European Community Innovation Survey (CIS) and the annual Mannheim Innovation Panel (MIP), which also constitutes the German contribution to the CIS. Both surveys provide firm-level information about innovative and non-innovative enterprises as well as R&D expenditure. Furthermore, they characterize an innovation by its degree of novelty[1] and the type of innovation[2] (Eurostat and OECD 2005). However, such indicators suffer from some major drawbacks. The MIP, for example, covers 10,000 firms every year, which corresponds to only 0.3% of the total number of firms in Germany. Thus, the total number of innovative firms remains unknown and can merely be estimated through statistical analysis. Furthermore, rare but potentially important innovation activities may not be covered in the data at all. This also affects the analysis of spatial processes within the STI system, some of which happen to operate on a fine (micro-)geographical scale (Carlino and Kerr 2015; Kerr et al. 2014; Arzaghi and Henderson 2008; Jang, Kim, and von Zedtwitz 2017; Catalini 2012). The effect of the presence of a university on the innovation activities of nearby firms, for example, is difficult to analyze if the local data sample of firms around the university is sparse. Consequently, established innovation indicators from questionnaire-based surveys lack granularity. Additionally, questionnaire-based surveys – especially on a large scale – are costly and time intensive. They also lack timeliness as it takes time to collect and process the data. Furthermore, surveys require firm participation as the questionnaire has to be answered by the firm. As a result, voluntary surveys like the MIP suffer from uncompleted questionnaires and the desired information is not always accessible (Kleinknecht, Van Montfort, and Brouwer 2002).

As an alternative to questionnaire-based surveys, innovation activity can be studied by analyzing patents (patent applications, citations, licensing). However, indicators constructed from patents cover only technological progress for which legal protection has been sought (Archibugi and Pianta 1996). Moreover, most patents are never used (Shepherd and Shepherd 2003); thus, they serve rather as indicators of inventions than of innovations. Another drawback of patent-based indictors derived from patent statistics, especially if they take a more selective approach, is that the dataset suffers from insufficient timeliness (Squicciarini and

[1] Innovations that are new to the firm, new to the market, new to the industry, or new to the world.

[2] Product innovations, process innovations, marketing innovations, and organisational innovations.

Criscuolo 2013). The time lag between priority date and the information becoming available is usually more than a year (OECD 2009). Literature-based innovation output indicators (LBIO) are constructed by counting innovations in scientific, technical, or trade journals. This indicator type is usually used to measure the degree of radicalness of innovations. However, LBIOs do not capture in-house process innovations and the measure can be inflated for some technologies which might help firm profits to improve by signaling innovativeness (Coombs 1996) or if other diverging incentives for firms to publish product innovations exist (Kleinknecht and Reijnen 1993). In addition, Acs, Anselin, and Varga (2002) indicate that LBIOs under-represent innovations in smaller firms as their influence on media is usually smaller.

We identified the following shortcomings which apply to a varying degree to traditional innovation indicators:

- *Coverage*: They cover only a fraction of the overall firm population.
- *Granularity*: They suffer from insufficient sectoral and technological granularity.
- *Timeliness*: They depict the state of the STI system as it was months or even years previously.
- *Cost*: They involve high data collection costs, especially when conducted on a large scale.

2.2. A general analysis framework for generating web-based innovation indicators

Nowadays, almost all (relevant) firms have their own websites which they use to publish information about their products and services. We assume that they also use this platform to highlight new and innovative features. In addition, firm websites provide additional information about firm credibility, achievements, key personnel decisions, strategies and relationships with other firms (Gök, Waterworth, and Shapira 2015). These aspects can all be related to a firm's innovation activity. Therefore, firm websites may reveal directly or indirectly whether new products, technologies, and processes are being implemented. While this data is publicly available, it is unstructured and stored in a decentralized manner. Therefore, there is a need for a consistent methodology for gathering and harmonizing the data, as well as for extracting innovation-related information.

In Figure 1, we outline just such a methodology in the form of a general analysis framework for generating web-based firm-level innovation indicators. Similar to traditional innovation indicators, the base data is a firm (panel-)database which includes information on firm characteristics (e.g. sector, firm size) and, most importantly, the firms' website addresses

(URLs). Ideally, the firm database has been matched to databases of established innovation indicators from questionnaire-based surveys, firm-level patenting data or literature data (LBIO), such that traditional indicators of innovation are available for a subsample of the firms. In a first step, the firms' web addresses are passed to a web scraper. The web scraper is then used to download website content (texts, hyperlinks etc.) from the firms' websites. In a third step, data mining techniques are applied to extract information on the firms' innovation activities from the downloaded website content. Based on this information, novel innovation indicators can be constructed. At this stage, additional metadata on the firm can be used to support the analysis (pre-classification, classification model selection based on firm characteristics, information from established innovation indicators etc.). In a final step, the newly generated innovation indicators are merged back into the firm database. This last step also establishes a direct firm-level link between the novel innovation indicator and the established indicators available from the auxiliary databases. This link can later be used to evaluate the new indicators against the traditional ones.

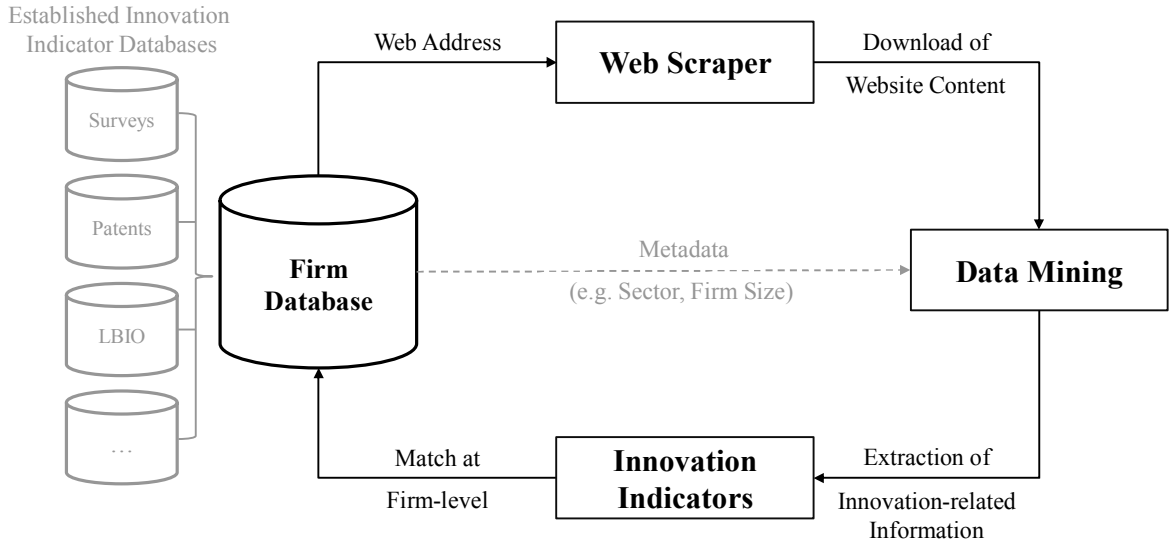


Figure 1. General analysis framework for generating web-based innovation indicators.

The proposed analysis framework allows for an automated, less costly analysis of entire firm populations than can carried out faster and in shorter intervals in comparison to traditional indicators. Furthermore, receiving firm information from websites does not require any effort on the part of the analyzed firms. As a result, web-based indicators have the potential to outperform traditional indicators in terms of coverage, granularity, timeliness, and survey costs. The crucial point in our proposed framework is the identification and extraction of those pieces of information from the unstructured website content that reveal information

about firms' innovation activities. We think that recent technological advances in analyzing unstructured data are making this possible (Steiger, Resch, and Zipf 2016; Mikolov et al. 2011; Grentzkow, Kelly, and Taddy 2017). Methods such as deep neural networks for natural language processing and social network analysis are able to deal with the difficulties resulting from heterogeneous data sources and to extract interpretable and meaningful information (see Conclusion and Future Research Section).

3. Previous Research

There are only a few existing studies analyzing the usability of web-based innovation indicators. These studies either employ web content mining or web structure mining (Miner et al. 2012). The latter is the analysis of connections between entities (e.g. firms) via the hyperlink structure of websites. Katz and Cothey (2006) used this approach to develop a method that produces indicators for the web presence of innovation systems. In a case study on European and Canadian education institutions, they find that their method is suitable for measuring "the amount of recognition a nation or province's web presence receives from other nations and provinces in their innovation systems" (Katz and Cothey 2006, 85). The authors emphasize the importance of reproducible and accurate indicators which are capable of dealing with the constantly changing properties of the Internet. Ackland et al. (2010) combine a web structure with a web content analysis.

In web content analysis, texts and other website content are analyzed. This approach is taken by the following studies: Youtie et al. (2012) use web scraping to explore the transitions from discovery to commercialization of 30 nanotechnology SMEs. Arora et al. (2013) use a similar approach to analyze entry strategies of SMEs commercializing emerging graphene technologies. Both study approaches are able to identify different innovation stages. Applying a keyword technique to explore the R&D activities of 296 UK-based enterprises, Gök, Waterworth, and Shapira (2015) find that web-based indicators offer additional insights when compared with patent and literature-based indicators. In addition, they emphasize that web mining as a research method has another advantage. The act of surveying a subject using web scraping does not cause certain problems such as altering the behavior of the study subject in response to being studied. The authors conclude "…that web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources" (Gök, Waterworth and Shapira 2015, 653). However, they raise the criticism that obtaining information from website data is more difficult and care needs to be taken when generating web-based indicators. The information on websites is

generally more related to innovation output than input. In addition, websites are self-reported and firms are not publishing new information on their websites at equal rates. Beaudry, Héroux-Vaillancourt and Rietsch (2016) use a keyword technique to generate innovation indicators of Canadian aeronautic, space and defense, as well as nanotechnology-related firms based on the text on their websites. They find some significant correlation between their indicators and traditional ones. Nathan and Rosso (2017) combine UK administrative microdata, media and website content to develop experimental measures of firm innovation for SMEs. The authors use proprietary data gathered by a data firm which uses website and media content to model firms' lifecycle events such as new product and service launches. They are able to identify three times more product/service launches than patent applications from SMEs in 2014/2015. Nathan and Rosso (2017) conclude that web-based indicators are a useful complementary measure to existing metrics as they reveal additional information. Moreover, they find that past patent activities are related to a firm's current launch activities and that tech SMEs are substantially more launch-active than non-tech SMEs.

The study by Kim et al. (2012) is also worth mentioning here. They do not make use of firm websites but apply text mining methods to forecast technology developments. The use data from published papers and patents to detect emerging technologies and determine their stage of development. As patents tend to detect inventions rather than innovations, firm websites promise to provide additional insights for measuring technology developments with text mining tools.

Studies on web-based innovation indicators have thus confirmed that firm websites are an interesting and rich data source for examining the innovation activity of firms and STI systems in general. However, no consistent approach (like the one we presented in the previous section) on how to study firms' websites has yet been established. Moreover, the data source itself (i.e. the population of firm websites) has not been studied rigorously in terms of its qualitative and quantitative properties. A number of basic yet important data characteristics are still unknown:

- *Structure*: Structural properties (size/depth, type of information provided, technological framework, web technologies used, update frequencies, languages used) of firm websites are largely unknown.
- *Coverage*: Coverage and structure of firm websites may differ systematically depending on the sector, firm size, firm age or region.

In the remainder of this paper, we attempt to fill in some of these gaps in knowledge by conducting a large-scale study using the whole population of German firm websites and additional information on the firms.

4. Data

The *Mannheim Enterprise Panel* (MUP) is a panel database which covers the total population of firms located in Germany. It contains about three million firm observations which are updated on a semi-annual basis. The data covers firm characteristics such as the industrial branch (NACE codes; a classification of economic activities in the European Union), postal addresses, number of employees, as well as the website address (URL) of the firm. For more information on the MUP see Bersch et al. (2014).

5. Methods

Note on terminology: A *website* is the overall internet presence of a firm. A website consists of a number of *webpages* (e.g. "www.firm-name.com", "www.firm-name.com/products"). The highest level webpage is called the *homepage* or the *main page* (e.g. "www.firm-name.com"), while lower level webpages are called *subpages* (e.g. "www.firm-name.com/products"), if a distinction has to be made. The first webpage downloaded from a website (the webpage corresponding to a URL in the user given list of URLs; this is usually the website's homepage) is referred to as the *start page*.

5.1. ARGUS web scraper

ARGUS (Kinne 2018) is an easy and free-to-use web scraping tool. The program is based on the Scrapy Python framework (Scrapy Community 2008) and is able to crawl a broad range of different websites to extract content like texts and hyperlinks. Full documentation on the program can be found in the appendix and online. An ARGUS web crawl is based on a list of firm website addresses (URLs) provided by the user and proceeds as follows:

1. The first webpage (usually a website's main page) is requested using the first address in the given URL list.
2. A collector item is instantiated, which is used to store the website's extracted content, meta-data (e.g. timestamps, number of scraped URLs etc.), and a so-called URL stack.

3. The main page is processed:
    a) Content from the main page is extracted and stored in the collector item.
    b) URLs which refer to subpages of the same website (i.e. domain) are extracted and stored in the collector item's URL stack.
4. The algorithm continues to request subpages of the website using URLs from the URL stack. To do this, it can use a simple heuristic which gives higher priority to short URLs and those URLs which refer to subpages in a predefined language.
    a) Content and URLs are collected from the subpage and stored in the collector item.
    b) The next URL in the URL stack is processed.
5. The algorithm stops processing a website once all subpages have been processed or as soon as a predefined number of webpages per website have been processed.
6. The collected content is processed and exported into an output file.
7. The next website is processed by requesting the next URL from the URL list. The described process continues until all firm website addresses from the list provided by the user have been processed.

Currently, ARGUS is able to scrape two kinds of web content: texts and hyperlinks. Figure 2 shows an exemplary extract of the database generated when using the scraper's option to download texts. In Figure 2, each line equals one webpage and *n* webpages equal one website. The number of downloaded webpages *n* per website is set by a *limit* parameter which can be adjusted by the user. If requested by the user, ARGUS uses a simple heuristic to select the subpages to download next after the initial landing webpage (*dl_rank*=0 in Figure 2) has been processed. Instead of requesting subpages in the order in which their URL appears in the website's HTML code, ARGUS continues with the shortest URL it finds. This may be a useful option if one is interested in more general information on the firm whose website is being scraped. We assume such general information ise usually located at the top level of websites (e.g. "firm-name.com/products", "firmname.com/team"). Additionally, ARGUS offers the option to preferentially select those URLs which indicate that they refer to a certain language. This is done by searching for ISO-3166 codes in the found URLs (e.g. "/de/" for German, "/fr/" for French). Our results show that this simple language heuristic helps to restrict the scraped texts to a certain language (see Results section).

	ID	dl_rank	dl_slot	error	redirect	start_page	text	timestamp	url
1471	313455	0	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/
1472	313455	1	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/
1473	313455	2	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/rss/
1474	313455	3	zew.de	None	False	https://www.zew.de/	Das Zentrum für Europäische Wirtschaftsforschu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/team/
1475	313455	4	zew.de	None	False	https://www.zew.de/	Unser informiert Sie direkt über Neuigkeiten a...	Wed May 23 13:06:30 2018	https://www.zew.de/de/presse/
1476	313455	5	zew.de	None	False	https://www.zew.de/	Tel: 0621 1235-132 Fax: 0621 1235-255 E-Mail: ...	Wed May 23 13:06:30 2018	https://www.zew.de/de/kontakt/
1477	313455	6	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/team/ybr
1478	313455	7	zew.de	None	False	https://www.zew.de/	Das ZEW ist ein gemeinnütziges wirtschaftswiss...	Wed May 23 13:06:30 2018	https://www.zew.de/de/das-zew/
1479	313455	8	zew.de	None	False	https://www.zew.de/	Wenn Sie den Hauptbahnhof verlassen haben, übe...	Wed May 23 13:06:30 2018	https://www.zew.de/de/anfahrt/

Figure 2. Exemplary extract of the database containing the scraped website texts of a single firm.

Figure 3 shows an exemplary extract of the database generated when using the scraper's option to download hyperlinks. Here, each line equals one website. Hyperlinks found on *n* ($n \leq limit$ parameter) webpages of each website are aggregated at the domain level (e.g. "firm-name.com"). While crawling a website, the algorithm proceeds with selecting the next sub-page URL to download as described above (i.e. URL selection heuristic can be used). The hyperlinks found on a website are split into two groups: hyperlinks to in-sample websites and hyperlinks to out-of-sample websites. *Links_internal* (see Figure 3) contains hyperlinks to only those websites which were in the initial list of firm website addresses provided by the user, while the *links_external* column contains both hyperlinks to in-sample websites and to other (out-of-sample) websites. This allows the user to analyze links between a set of predefined firms (e.g. only German firms), as well as their wider (e.g. non-German) network, separately. Note, however, that out-of-sample websites are not crawled and ARGUS, thus, does not collect information on whether connections to out-of-sample website are reciprocal.

ID	alias	dl_slot	error	links_internal	links_external	redirect	timestamp	url
3	NaN	bmbf.de	None	bmbf.de,ec.europa.eu,leibniz-gemeinschaft.de	bmbf.de,bmbf-besuchen.de,twitter.com,facebook....	False	Thu Jul 26 10:44:19 2018	https://www.bmbf.de/
1	NaN	zew.de	None	zew.de,uni-mannheim.de,leibniz-gemeinschaft.de...	zew.de,seek.zew.eu,kooperationen.zew.de,wgl.de...	False	Thu Jul 26 10:50:30 2018	https://www.zew.de/
4	NaN	uni-mannheim.de	DNS	NaN	NaN	NaN	Thu Jul 26 10:41:00 2018	NaN
5	NaN	leibniz-gemeinschaft.de	None	leibniz-gemeinschaft.de,bmbf.de,ec.europa.eu	leibniz-gemeinschaft.de,twitter.com,facebook.c...	False	Thu Jul 26 11:34:05 2018	https://www.leibniz-gemeinschaft.de/start/

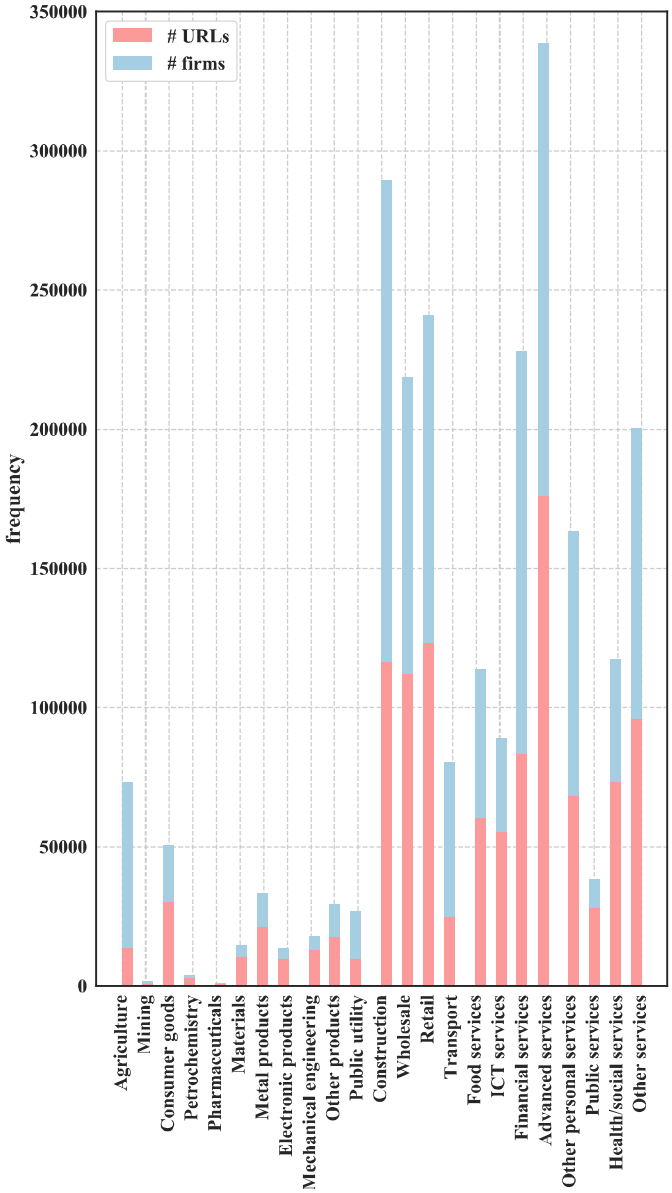
Figure 3. Exemplary extract of the database containing the scraped hyperlinks texts of a five different firms.

6. Results

6.1. URL coverage

For our analysis, we use Mannheim Enterprise Panel data from early 2018 and restrict our data to firms which were definitely economically active at that time. This leaves us with about 2.52 million firms and URLs for about 1.15 million firms (46% coverage) (see Table 1). URL coverage differs by sector, firm size (in terms of employees), and the firms' age. Table 1 breaks down the firm population by sector (variable available for 96% of firms; sector classification in Table A1 in the appendix). From the table, it can be observed that some sectors are fairly well covered (≥70% coverage for materials, electronic products, mechanical engineering, and public services), while other sectors are poorly covered in our dataset (≤40% coverage for agriculture, public utility, construction, transport, financial services).

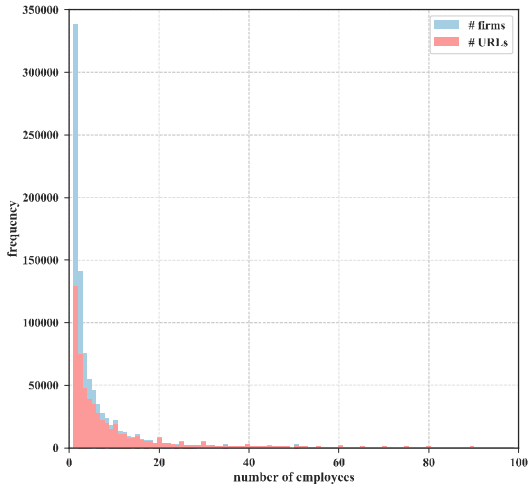
Table 1. Number of firms and number of available URLs by sector.



Sector	# firms	# URLs	:
Agriculture	73,111	13,507	0.18
Mining	1,528	795	0.52
Consumer goods	50,423	30,130	0.60
Petrochemistry	3,617	2,489	0.69
Pharmaceuticals	928	594	0.64
Materials	14,628	10,218	0.70
Metal products	33,267	20,934	0.63
Electronic products	13,432	9,675	0.72
Mechanical engin.	17,781	12,677	0.71
Other products	29,267	17,600	0.60
Public utility	27,038	9,718	0.36
Construction	289,399	116,137	0.40
Wholesale	218,664	112,070	0.51
Retail	240,857	123,141	0.51
Transport	80,373	24,585	0.31
Food services	113,688	60,258	0.53
ICT services	89,061	55,062	0.62
Financial services	227,927	83,064	0.36
Advanced services	338,519	175,774	0.52
Other personal serv.	163,391	68,133	0.42
Public services	38,217	27,918	0.73
Health/social serv.	117,383	73,084	0.62
Other services	200,293	95,896	0.48
MISSING sector	140,439	5,389	0.04
Total	**2,523,231**	**1,148,848**	**0.46**

Table 2 shows the URL coverage by the number of employees firms have (variable available for 38% of firms). We can see that most firms are very small (micro-enterprises with less than 6 employees) and that coverage for this group is rather low (49%). For small firms (6-25 employees) coverage is decent (84%). Medium (26-250 employees) and large firms (>250 employees) are covered very well (94%; 97%). These numbers are in line with official statistics, which cite the share of enterprises in Germany with websites at 87% for firms with 10 or more employees and 64% for firms with less than 10 employees (Eurostat 2018). A two-sample t-test (see e.g. Krzywinski and Altman 2013) indicated a highly significant difference in the number of employees between the overall firm population ($\bar{x}$=3.4) and the subpopulation covered by a URL ($\bar{x}$=19.6).

Table 2. URL coverage by firms' number of employees.

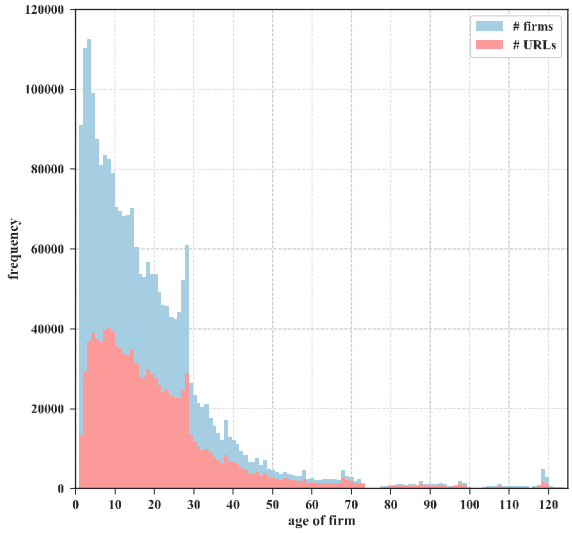


# employees	# firms	# URLs	:
1-5	655,617	324,393	0.49
6-25	229,995	193,648	0.84
26-250	71,778	67,132	0.94
>250	6,481	6,298	0.97
all	**963,871**	**591,471**	**0.61**

Table 3 shows the URL coverage by firm age (variable available for 91% of firms). Several historical events which led to an increased firms being founded can be seen in the firms' age distribution (left): German Reunification (~28 years), constitution of the Federal Republic after the Second World War (~70 years), and the entrepreneurial boom of the *Gründerzeit* (~120 years). Furthermore, URL coverage in the data increases with firm age. While very young firms (younger than two years) are poorly covered (18%), firms which are older than six years have better coverage (about 50%). It should be noted that firm age and firm size are highly correlated (Spearman's rho of 0.37; $p<0.001$). A two-sample t-test indicated a highly significant difference in the age the overall firm population ($\bar{x}$=16.7) and the covered subpopulation ($\bar{x}$=21.2).

Figure 4 maps the ratio of firms with an available URL to the overall firm population by district. We can observe that low and high ratios do not seem to be randomly scattered, but instead low coverage can be primarily found in East Germany, while West Germany seems to be well covered. This impression of non-randomness is confirmed by a high and significant *Moran's I* (see e.g. Fischer and Getis 2010) value of 0.39 ($p<0.001$) indicating high positive spatial autocorrelation (clustering). We further identified several significant ($p<0.05$) local clusters of both high (West and South-West) and low (East and North-East) URL coverage using *Getis-Ord Gi** (Getis 2009) measure of local autocorrelation. We also find that coverage is generally better in densely populated (urban) areas, indicated by a very high and significant correlation between population density and URL coverage (Spearman rho of 0.5; $p<0.001$). This relation can also be seen in Figure 4.

Table 3. URL coverage by firms' age.



Age	# firms	# URLs	:
0-1	138,922	25,005	0.18
2-5	395,725	146,670	0.37
6-25	1,210,762	616,722	0.51
26-100	521,821	273,435	0.52
>100	36,138	20,548	0.57
all	**2,303,368**	**1,082,380**	**0.47**

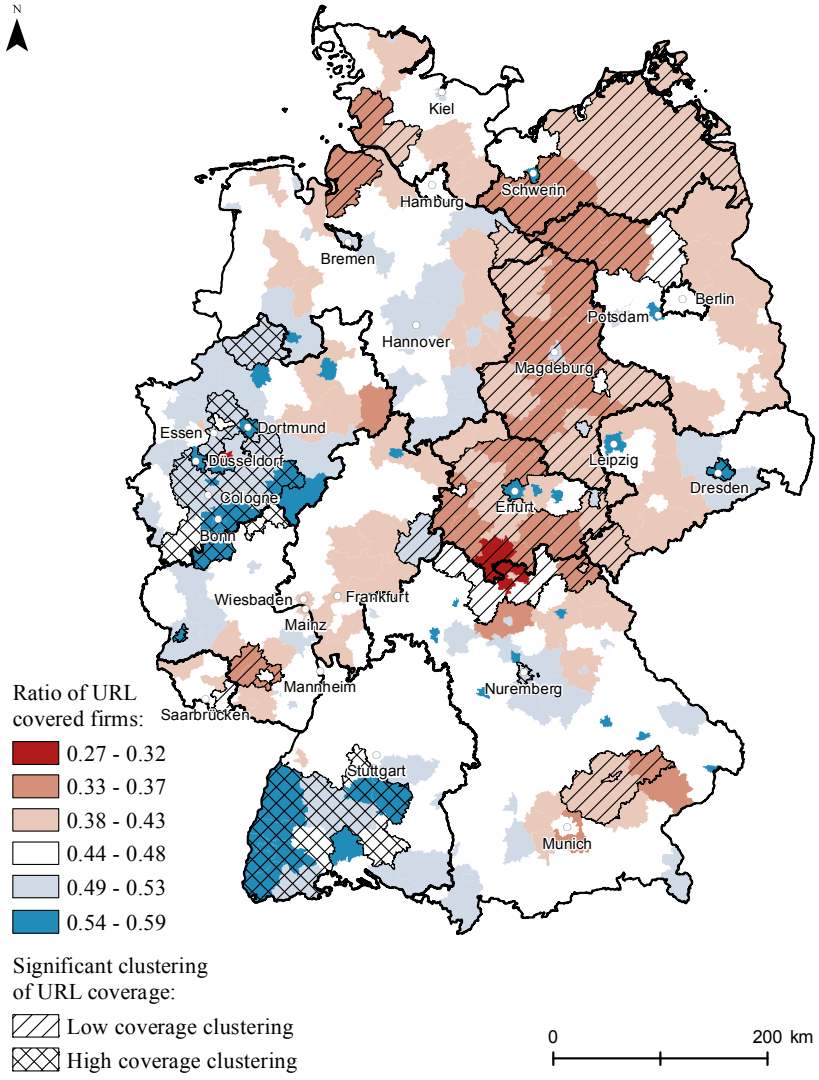


Figure 4. Ratio of URL covered firms by districts.

Missing URLs in our data can result from either incomplete inquiry by the data provider or the fact that firms have no website. We control for this using two control variables. Some legal forms require a mandatory entry in official commercial registries – a procedure which makes surveying the firm a lot easier and, thus, increases the probability of a correctly entered URL in our data. We use information on the firms' *legal form* to control for this. The *search quality* variable controls for bias in the data provider's search strategy as well. We use the availability of a phone number in our data as an indicator for how well the firm was researched by the data provider.

Table 4 shows the results (marginal effects) of a probit regression with URL availability in our data as the dependent variable. *Broadband availability* is measured as the percentage of households in the firm's municipality that have potential access to broadband internet (≥50 Mbits download speed available; all technologies) (BKG, BMVI, and TÜV Rheinland 2016). *Population density* controls for urban or rural firm locations and makes sure that broadband availability is not just a proxy for urban/rural firm location. *Employees*, *age*, and *sector* are defined as above. Our baseline firm is a mechanical engineering firm in a region with >95% broadband availability, 0 population density (rural area), >250 employees, >100 years of age, a legal form which requires an entry in the German commercial registry, and with an available phone number in our data. The pseudo-R^2 of the model is 0.19 and the mean variance inflation factor is 9.36, indicating unproblematic multicollinearity.

Our findings above are confirmed by the probit regression. Very young and very small firms do not have websites, and the firm's sector plays a vital role as well. The regression also shows that firms in areas with low broadband availability are less likely to have a website. Our controls make us confident that this is not just a bias in the search strategy of our data provider. Instead, low broadband availability may detain firms from going digital and running their own website. According to the estimated effects, 30,000 firms (extrapolated to the total firm population) do not have a websites because of their region's low high-speed Internet availability. This relates to 3.6% of firms in poor Internet regions, and to 1% of the total firm population in Germany respectively.

Table 4. Probit regression results: Firm has/has no (1/0) URL.

Variable	Marginal effect	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	-0.001	0.001
50-75%	-0.022***	0.001
10-50%	-0.044***	0.001
0-10%	-0.057***	0.002
Population density		
1,000 people/km²	0.008***	0.000
Employees		
MISSING	-0.484***	0.005
1-5	-0.373***	0.005
6-25	-0.134***	0.005
26-250	-0.041***	0.006
Age		
0-1	-0.242***	0.003
2-5	-0.093***	0.003
6-25	-0.061***	0.003
26-100	-0.072***	0.003
Sector		
Agriculture	-0.308***	0.004
Mining	-0.188***	0.013
Consumer goods	-0.052***	0.004
Petrochemistry	0.014	0.009
Pharmaceuticals	-0.027	0.016
Materials	-0.010	0.005
Metal products	-0.075***	0.005
Electronic products	0.030***	0.006
Other products	-0.041***	0.005
Public utility	-0.201***	0.005
Construction	-0.197***	0.004
Wholesale	-0.095***	0.004
Retail	-0.077***	0.004
Transport	-0.282***	0.004
Food services	-0.040***	0.004
ICT services	0.053***	0.004
Financial services	-0.176***	0.004
Advanced services	-0.060***	0.004
Other personal services	-0.129***	0.004
Public services	0.136***	0.004
Health/social services	0.021***	0.004
Other services	-0.030***	0.004
Legal form		
Registry entry not mandatory	-0.059***	0.001
Foreign legal form	0.358***	0.020
Search quality		
No other contact info	-0.362***	0.001

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old, has legal form which requires entry in commercial registry, and other contact info (phone) is available in data.

*p≤0.05, **p≤0.01, ***p≤0.001; n=2,108,104

6.1.1. Request errors and URL redirects

We randomly draw 11,477 firms for which an URL is available and used ARGUS to scrape their websites. Doing so, we were able to scrape 84.2% of the websites. The remaining 15.8% websites returned errors (DNS errors, timeouts, and HTTP errors) when requesting their start pages. T-tests between firms with successfully/not successfully requested websites showed no significant difference in firm size and age. However, some sectors (e.g. mechanical engineering, consumer goods, and petrochemistry) had fewer errors than others (financial services, public services) (see Table 5).

Table 5. Request errors by sector.

Sector	# firms	# errors	:
Agriculture	107	12	0.11
Mining	7	1	0.14
Consumer goods	300	31	0.10
Petrochemistry	18	2	0.11
Pharmaceuticals	6	1	0.17
Materials	105	13	0.12
Metal products	204	27	0.13
Electronic products	109	14	0.13
Mechanical engineering	134	11	0.08
Other products	165	19	0.12
Public utility	107	18	0.17
Construction	1,218	156	0.13
Wholesale	1,147	180	0.16
Retail	1,217	223	0.18
Transport	222	28	0.13
Food services	608	110	0.18
ICT services	557	82	0.15
Financial services	836	224	0.27
Advanced services	1,718	258	0.15
Other personal services	691	104	0.15
Public services	280	53	0.19
Health/social services	702	100	0.14
Other services	963	138	0.14
Total	**11,421**	**1,805**	**0.16**

We further investigated the share of URLs for which initial requests are redirected. We only tag redirects if the redirect results in crawling a webpage from a different (second level) domain (e.g. “www.example.com” redirects to “www.sample.com”). Redirects between secure and standard HTTP (e.g. “http://www.example.com” to “https://www.example.com”) and subdomain changes (e.g. “www.products.example.com” to “www.example.com”) are

not tagged as redirects. Redirects we tag can be both harmless (e.g. a firm registered a new domain and redirects there from its old domain) and severe (e.g. firm A was acquired by firm B and firm A's old URL now redirect to website of parent company B; small firms sometimes register domains but redirect to personal pages on social media like facebook.com). To be sure that the crawled website really belongs to the corresponding firm, redirected requests must either be checked thoroughly or excluded from the analysis. We opt for the latter. Overall, 9,5% of URLs we successfully requested redirected. T-tests showed no significant difference in firms' age and size between redirecting and non-redirecting URLs. Again, some sectoral differences can be seen in Table 6.

Table 6. URL redirections by sector.

Sector	# firms	# redirects	:
Agriculture	95	5	0.05
Mining	6	0	0.00
Consumer goods	269	29	0.11
Petrochemistry	16	2	0.13
Pharmaceuticals	5	0	0.00
Materials	92	7	0.08
Metal products	177	21	0.12
Electronic products	95	14	0.15
Mechanical engineering	123	12	0.10
Other products	146	14	0.10
Public utility	89	12	0.13
Construction	1,062	60	0.06
Wholesale	967	111	0.11
Retail	994	100	0.10
Transport	194	17	0.09
Food services	498	32	0.06
ICT services	475	55	0.12
Financial services	612	75	0.12
Advanced services	1,460	157	0.11
Other personal services	587	49	0.08
Public services	227	26	0.11
Health/social services	602	48	0.08
Other services	825	67	0.08
Total	**9,616**	**913**	**0.09**

6.2. Website characteristics

6.2.1. Number of subpages

Excluding websites which resulted in request errors or redirects reduces our firm sample by 23.8% to 8,744 firms. For the subsequent analysis, the ARGUS web scraper *limit* parameter, which defines the maximum number of subpages scraped from a single website, was set at 2,500. The mean number of webpages per website is 218.8 (SD 604.7) and the median is 15, resulting in a highly skewed (3.24) distribution, as can also be seen in Figure 5. A considerable share (5.86%) of the websites reached the crawler's limit of 2,500 subpages, indicating that an unlimited crawl would result in an even more skewed distribution.

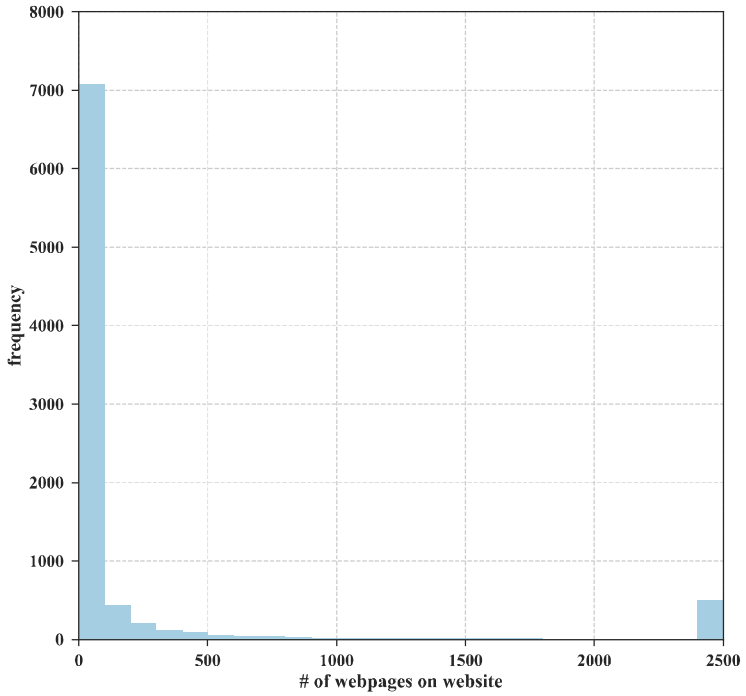


Figure 5. Histogram of number of webpages on firms' websites.

Figure 6 shows that the median number of webpages per website does not differ considerably between sectors. The mean number of webpages does so though. A regression analysis (Table A2 in the appendix) indicates that only certain sectors (pharmaceuticals, retail, ICT services, financial services, and other services) have a significant effect on the number of webpages when compared against the baseline sector of mechanical engineering when controlling for the firms' size and age. However, the predictive performance of the regression model is rather low (R^2=0.06). Both the regression and a high correlation coefficient (Spearman's rho of 0.19; $p<0.001$) between firm size (number of employees) and the number of webpages indicate that larger firms have larger websites.

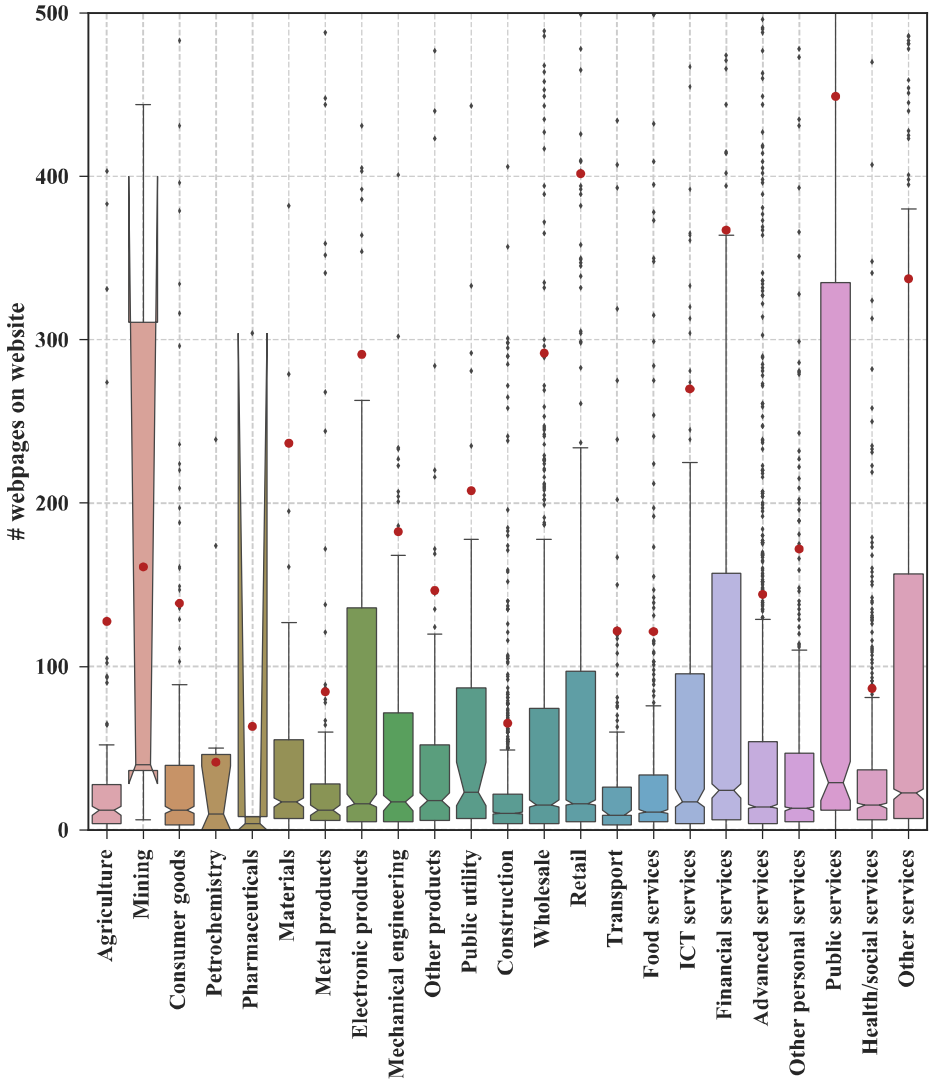


Figure 6. Notched boxplot of number of webpages on firms' websites by sector; means as red dots.

6.2.2. Text volume

On average, a webpage has 3295.86 characters (SD=9960.43) and half of the webpages have 1970.8 characters or less (about two thirds of a page of text in this paper), resulting in a highly skewed (39.05) distribution as seen in Figure 7.

An OLS regression analysis (Table A3 in the appendix) shows that the mean number of characters per webpage does not relate to a firm's characteristics (age, size, and sector). Consequently, the predictive performance of the model is very low (R^2=0.01).

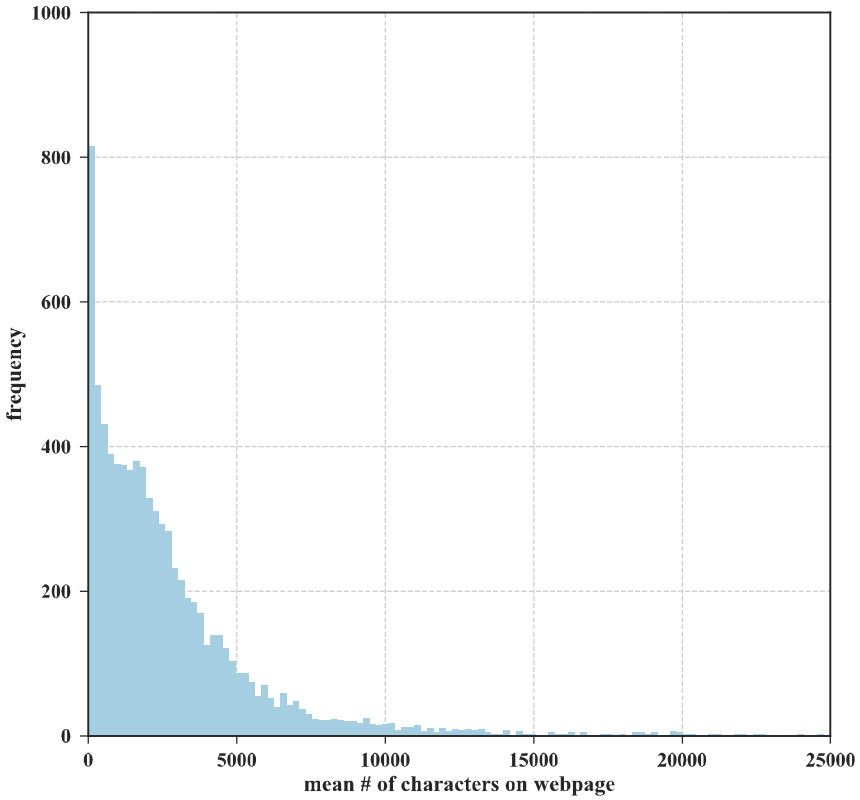


Figure 7. Histogram of mean number of characters on webpage.

6.2.3. Text language

We randomly drew 911 websites from our sample above and detected the languages used in all of their 193,504 webpages using Python's langdetect library (Danilak 2015). 91.9% of the webpages could be classified and of these 88.2% were classified as being written in German. Most (60.8%) of the non-German language webpages were classified as written in English. Manual checking indicated that the share of English language webpages is likely to be slightly overestimated due to the frequent use of English vocabulary in otherwise non-English text. Most websites are written almost completely in German (close to 100% of their webpages are classified as German), as can be seen in Figure 8. The regression analysis in Table A4 in the appendix shows that the share of German language on a firm's website is related to the firm's sector, while the firm's size, age, and location does not play a significant role. Compared to the baseline mechanical engineering firm, only firms from the pharmaceutical sector use less German language on their websites, while firms from other sectors (e.g. agriculture, mining, metal products, public services) use more German. The predictive performance of the model is rather high (R^2=0.17) too.

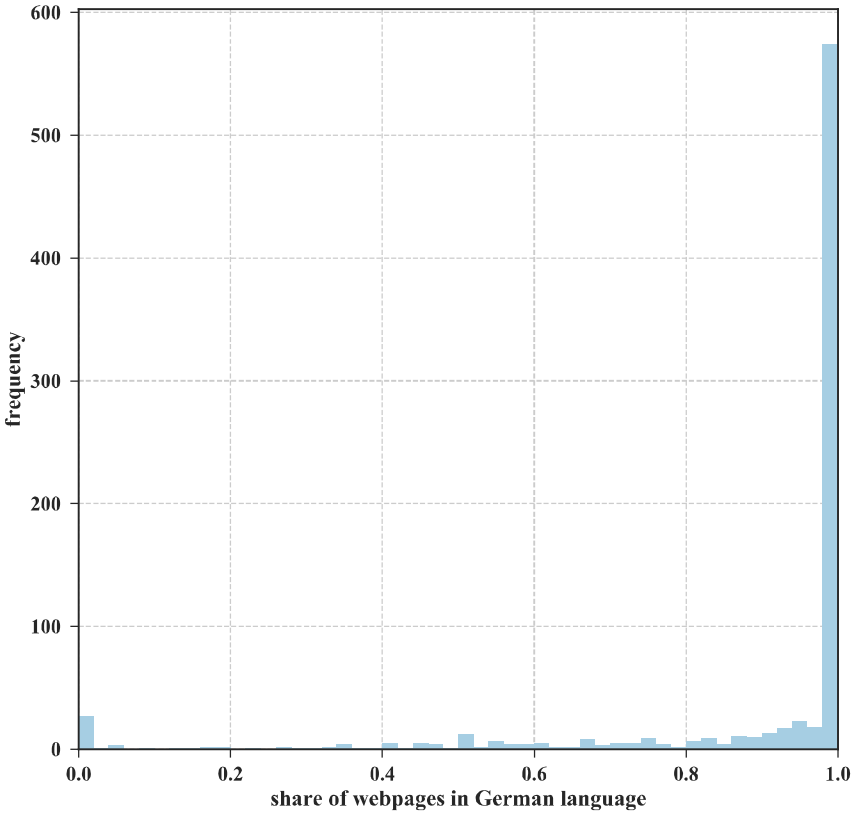


Figure 8. Histogram of share of webpage in German language.

It is important to keep in mind that the webpages were not selected uniformly or randomly from the firms' websites, as we used the ARGUS language selection heuristic set to German. Consequently, a high share of webpages with text in German does not necessarily imply that the firm exclusively uses German on its website. Changing the preferred language to English decreases the share of German classified webpages from 88.2% to just 74.9% and increases the share of English webpages from 7.2% to 11.3%. This indicates that some firms have a German and an English version of their website and ARGUS is able to scrape a preferred language if desired by the user – a desirable feature as most natural language processing methods require text corpora in a single language. The fact that some firms only have non-German texts on their websites (share < 0.2 in Figure 8; 4.5%) indicates that some Germany-based firms do not have a German version of their websites at all.

6.2.4. Hyperlinks

We scraped links from the websites of our initial firm sample of 8,744 observations (see 5.2.2; *limit* parameter set to 100). The resulting distribution of hyperlinks per website is extremely skewed (72.14), with no website having less than 14 hyperlinks and some outlier websites including tens of thousands of hyperlinks: The mean number of hyperlinks per website is 252.17 (SD 1779.69) and the median is 116. Unsurprisingly, the number of hyperlinks found on a firm's website is highly correlated (Spearman's rho of 0.51; $p<0.001$) with the website's overall size (i.e. number of webpages).

Hence, it is more interesting to investigate the mean number of hyperlinks per webpage. On average, a webpage contains 14.52 hyperlinks. The median number of hyperlinks per webpage is just 6, resulting in a highly skewed (30.63) distribution as seen in Figure 9. In Figure 10, we can see that the median is rather stable over sectors, while the mean fluctuates somewhat between sectors (driven by outlier firms). The OLS regression (Table A5 in the Appendix) has a very low predictive performance ($R^2=0.006$) and indicates that firm characteristics (sector, size, age, broadband availability) play no significant role in the mean number of links per webpage.

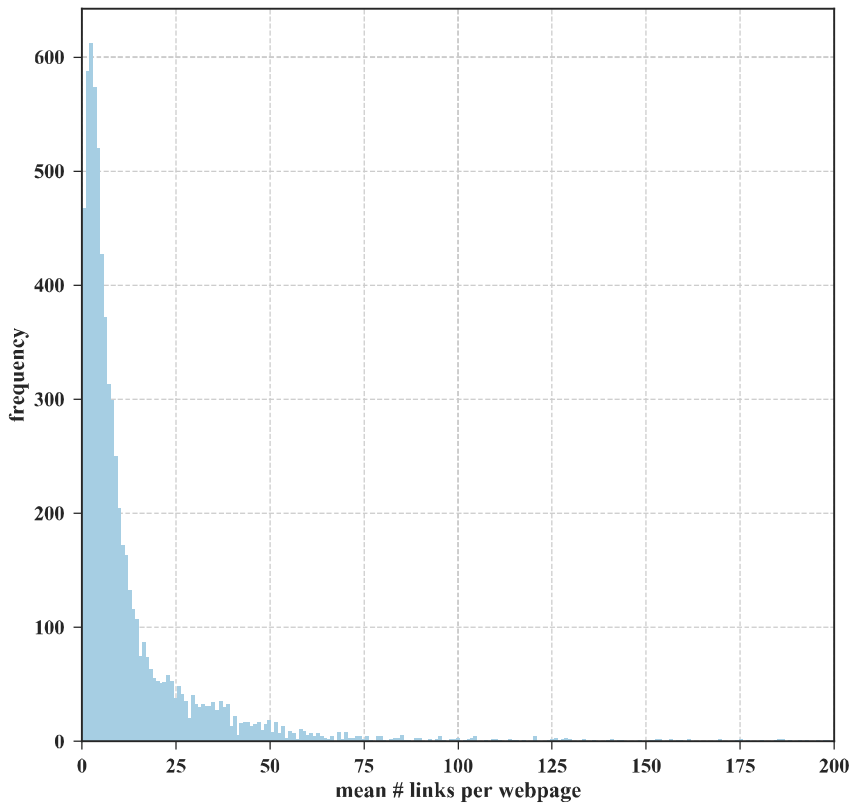


Figure 9. Histogram of number of links per website.

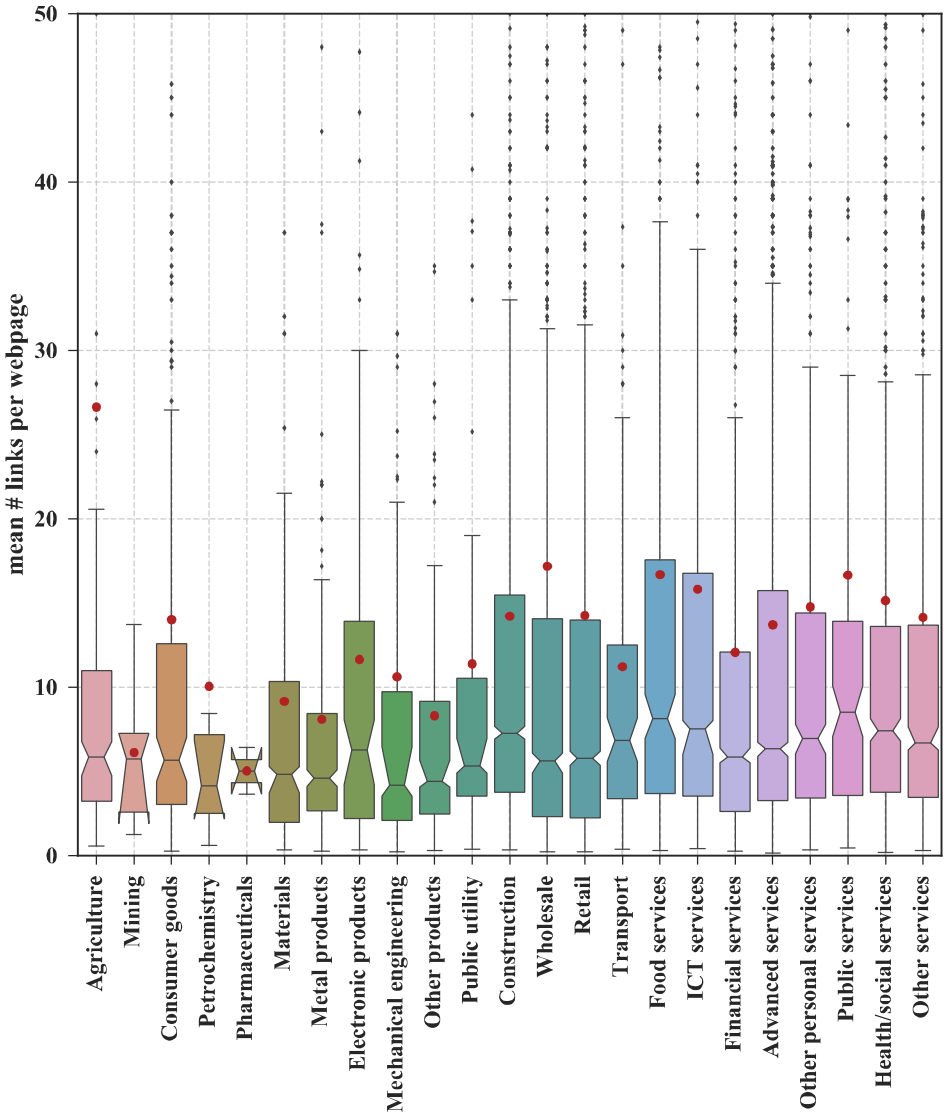


Figure 10. Notched boxplot of number of links on firms' websites by sector; means as red dots.

7. Discussion

We first analyzed the availability of firm website URLs in our dataset (URL coverage) and investigated regularities concerning URL coverage and firms' characteristics. We also tried to untangle the cause of missing URLs in our dataset and distinguish between *true* missing values (the firm has no website) and *false* missing values (the firm has a website, but it was not found by the data provider). Based on our case study results, regularities in the URL coverage remain even after controlling for bias in the search strategy of the data provider. Using our proposed framework for generating web-based indicators, there may be difficulties when observing very young and very small firms, especially from certain sectors such as agriculture and in rural areas. In addition, low broadband availability seems to hinder firms from going digital and setting up a website and therefore systematically excludes them from web-based studies. If one assumes that low broadband availability is associated with low private and commercial use of the internet, this may indicate that firms with more local markets and located in an area of low broadband availability have no incentive to set up their own website in order to communicate with customers. On the other hand, our results show that medium-sized and medium-aged, as well as large firms can be thoroughly surveyed using our proposed approach. This is especially true in urban areas. Given that such firms conduct the vast majority of innovative activity in Germany (C Rammer et al. 2017), we conclude that our approach and our data is suitable for analyzing the German STI system.

We identified URL redirects as a potential issue when conducting web-based studies. Out-dated URLs may result in potentially harmful redirects (see 5.1.1). If one is working with large-scale datasets, it is usually not possible to make sure that the available firm website addresses are all up-to-date. To ensure unbiased results, we recommend excluding firms with URL redirects from web-based studies. Given that less than 10% of successful URL requests were redirected, and we did not find any systematic firm age or size bias, excluding URL redirects seems reasonable. We also identified a non-uniform distribution of redirects across sectors, which may result from strong dynamics (firm start-ups, closured, mergers and acquisitions) within some sectors.

Our results showed that firm website size (i.e. number of webpages) is highly correlated to a firm's size (number of employees) as well as to a firm's sector. Large firms have larger websites (more webpages on their websites), but they do not provide (on average) more text on each of these webpages. In general, we find that outliers play an important role when conducting web-based studies. Some websites are extremely large in terms of the number of webpages and the amounts of text. This outlier issue also causes the mean number of

webpages per website to vary quiet strongly between sectors. On the other hand, the median number of webpages per website is rather stable across sectors (about 15 webpages per website). To fully scrape two thirds of all firm websites completely, it is sufficient to set the maximum webpage download per website limit to 50. Fully scraping 90% of the firms' websites requires that this threshold be raised to 250. About 6% of firms can be seen as extreme outliers with 2,500 or more webpages. Based on these purely quantitative results, it is difficult to make any best practice recommendation since only further qualitative investigations can reveal what webpage threshold is sufficient to scrape the desired information from firms' websites. However, our results should provide researchers with a sound reference point when conducting their own studies.

Concerning the language used on the firms' websites, our results showed that, unsurprisingly, most websites of Germany-based firms are in German. However, a considerable share (about 5%) of the firms have mostly (≥80%) non-German texts on their websites. Given that most natural language processing algorithms require text corpora to be in a single language, this is a significant result. We were also able to show that the ARGUS simple language selection heuristic helps to restrict the texts downloaded to a certain language. The same results also indicate that a considerable share of firms provide several versions of their website in different languages. This ARGUS language selection heuristic is likely to be even more important when working with websites from multilingual countries (e.g. Switzerland, Belgium). Furthermore, we found significant sectoral differences in the use of language. Some sectors (e.g. agriculture, personal services, construction) mostly use German, while others (e.g. mechanical engineering, pharmaceuticals) use other languages as well. We assume that the sector's orientation towards either local/national or international markets may play an important role here.

The total number of hyperlinks that can be found on a firm's websites is, unsurprisingly, highly correlated to the number of webpages it has. The mean number of links per website, however, seems to be randomly distributed with no significant relationship to the firm's size, age, or sector. If hyperlinks between firms are interpreted as some kind of relationship (e.g. customer, cooperation), this would indicate that, on average, the connectedness of a firm grows with its size. A qualitative analysis of these connections could reveal whether certain types of firms (e.g. innovative ones) are connected differently (e.g. in the degree of local, intra-sector connectedness) compared to other types (e.g. non-innovative firms).

8. Conclusion and Future Research

8.1. Conlusion

In this paper, we proposed a framework for generating indicators for the web mining of firms' websites and illustrated, using the example of innovation indicators, how these novel indicators can be a useful addition to the existing set of indicators available. We argued that established innovation indicators have a number of shortcomings concerning their coverage, granularity, timeliness, and associated costs and that web-based indicators have the potential to overcome these limitations. The proposed framework is composed of four key parts: a database with firm-level metadata and firms' web addresses, a web scraper to download firm website content, data mining to extract information from the downloaded content, and the actual innovation indicators generated from the extracted information. The remainder of the paper dealt with the first two, presenting a universally applicable web scraper and a pilot study to investigate the properties of German firm websites. During the pilot study, we tackled two research questions.

8.1.1 RS1: URL coverage

URL coverage differs systematically based on firms' characteristics and, thus, excludes certain firm types from web-mining surveys. Only a fraction of very young and very small firms can be observed using our proposed approach. We also find sectoral and regional differences. Some sectors and regions exhibit very low URL coverage, while others are thoroughly covered. Furthermore, we find that low local broadband availability can prevent firms from setting up their own internet presence. We further identified initial HTML redirects caused by out-of-date website address data as a potentially harmful error source. We recommend excluding all HTML redirects from further analysis.

8.1.1 RS2: Website characteristics

From our analysis results, we concluded that web-based studies have to deal with outlier issues, with about 6% of firms having websites with a number of webpages four or more standard deviations from the population mean. This issue is even more pronounced concerning the amount of text and hyperlinks found on firms' websites. Large firms not only operate larger websites, they also provide disproportionally more hyperlinks on these platforms. We also find that the number of webpages per website differs depending on the firms' sector, as does language, with some sectors making significantly greater use of non-German languages than others. We were also able to show that the ARGUS language selection heuristic helps

to restrict text downloads to a certain language and to exploit the fact that many firms provide several different language versions of their website.

8.1. Future Research

8.1.1 Estimating firms' innovation activities using neural networks

For future text analysis, we propose an approach for estimating a firm's innovation activity as outlined in Figure 11. A neural network is trained using texts scraped from websites of firms for which established innovation indicators are available. Such indicators can be used to create a training dataset of labelled (innovative/non-innovative) website texts. After training the neural network, unlabeled website texts (i.e. texts from websites of firms with unknown innovation activity) can be examined by the network and given a probability of being scraped from an innovative firm's website. Given that such information is available, additional firm metadata (e.g. the sector of the firm) can be used to enhance the model. Recent developments (e.g. Mikolov *et al.*, 2011, 2013; Mikolov, Yih and Zweig, 2013) in the field of natural language processing (NLP) make this approach potentially the most promising when it comes to inferring information about firms' innovation activity based on their websites' textual content.

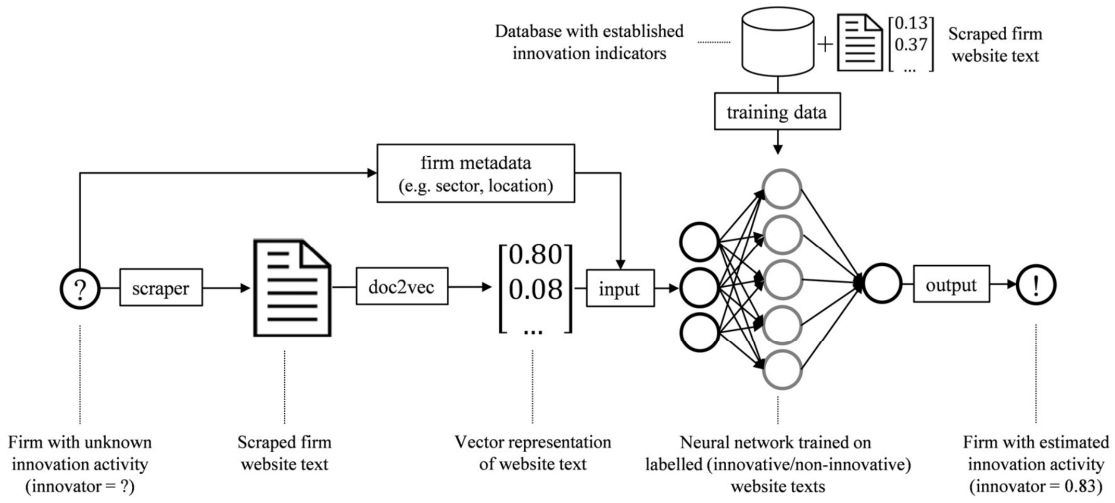


Figure 11. Approach for estimating a firm's innovation activity using a neural network.

8.1.2 Social network analysis as a means of hyperlink structure analysis

In this paper, we showed that networks of interconnected firms can be extracted from the web using ARGUS web scraper. Given that the appropriate metadata is available, specific regional and sectoral firm networks can be examined, like the one shown in Figure 12, which maps an exemplary network of software firms based in Berlin, Germany. Social network analysis offers an extensive set of widely adapted techniques for analyzing such networks in a quantitative manner (see e.g. Scott and Carrington 2011). Future research should aim to find regularities in the structure of firms' hyperlink networks, preferably by using established innovation indicators to differentiate between innovative and non-innovative firms and firm segmentations. Datasets like the one shown in Figure 12 could also be used to investigate the relatedness of firms on a microgeographic level of analysis, which is already an active string of research (see e.g. Christian Rammer, Kinne, and Blind 2016).

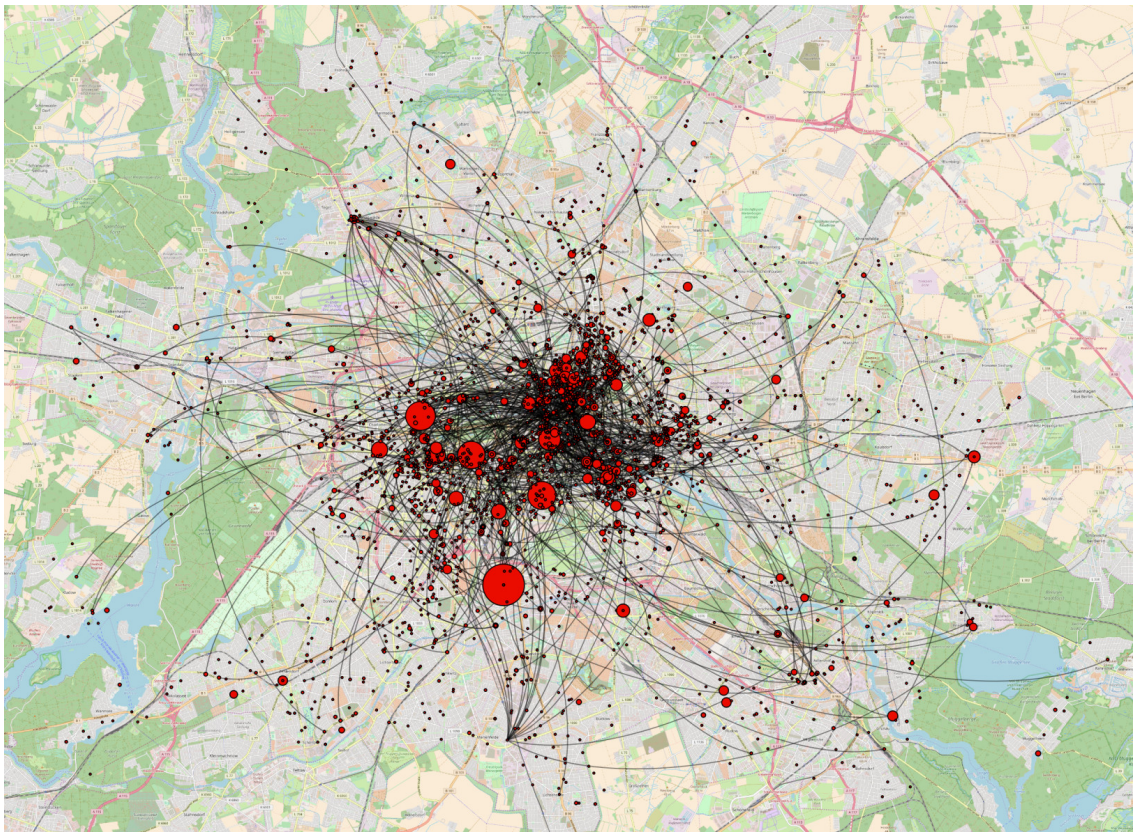


Figure 12. Exemplary map of hyperlink connections between software firms based in Berlin, Germany. Basemap: OpenStreetMap

References

Ackland, Robert, Rachel Gibson, Wainer Lusoli, and Stephen Ward. 2010. "Engaging With the Public? Assessing the Online Presence and Communication Practices of the Nanotechnology Industry." *Social Science Computer Review* 28 (4): 443–65.

Acs, Zoltan J., Luc Anselin, and Attila Varga. 2002. "Patents and Innovation Counts as Measures of Regional Production of New Knowledge." *Research Policy* 31 (7): 1069–85. https://doi.org/10.1016/S0048-7333(01)00184-6.

Archibugi, Daniele, and Mario Pianta. 1996. "Measuring Technological Change through Patents and Innovation Surveys." *Technovation* 16 (9): 451–68. https://doi.org/10.1016/0166-4972(96)00031-4.

Arora, Sanjay K., Jan Youtie, Philip Shapira, Lidan Gao, and TingTing Ma. 2013. "Entry Strategies in an Emerging Technology: A Pilot Web-Based Study on Graphene Firms." *Scientometrics* 95 (3): 1189–1207.

Arzaghi, Mohammad, and J. Vernon Henderson. 2008. "Networking off Madison Avenue." *Review of Economic Studies* 75 (4): 1011–38. https://doi.org/10.1111/j.1467-937X.2008.00499.x.

Askitas, Nikolaos, and Klaus F. Zimmermann. 2015. "The Internet as a Data Source for Advancement in Social Sciences." *International Journal of Manpower* 36 (1): 2–12. https://doi.org/10.1108/IJM-02-2015-0029.

Beaudry, Catherine, Mikaël Héroux-Vaillancourt, and Constant Rietsch. 2016. "Validation of a Web Mining Technique to Measure Innovation in High Technology Canadian Industries." In *CARMA 2016–1st International Conference on Advanced Research Methods and Analytics*, 1–25.

Bersch, Johannes, Sandra Gottschalk, Bettina Müller, and Michaela Niefert. 2014. "The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany." *ZEW Discussion Paper*. https://doi.org/10.2139/ssrn.2548385 M4 - Citavi.

BKG, BMVI, and TÜV Rheinland. 2016. "Broadband Atlas." Berlin. https://www.bmvi.de/DE/Themen/Digitales/Breitbandausbau/Breitbandatlas-Karte/start.html.

Carlino, Gerald, and William R. Kerr. 2015. "Agglomeration and Innovation." In *Handbook of Regional and Urban Economics*, edited by Gilles Duranton, J Vernon Henderson, and William C. Strange, 5:349–404. Amsterdam: Elsevier North-Holland. https://doi.org/10.1016/B978-0-444-59517-1.00006-4.

Catalini, Christian. 2012. "Microgeography and the Direction of Inventive Activity." *Rotman School of Management Working Paper*. Vol. 2126890. https://doi.org/10.1287/mnsc.2017.2798.

Coombs, Rod. 1996. "Core Competencies and the Strategic Management of R&D." *R&D Management* 26 (4): 345–55. https://doi.org/10.1111/j.1467-9310.1996.tb00970.x.

Danilak, Michal. 2015. “Langdetect.” https://pypi.org/project/langdetect/.

Eurostat. 2018. “EUROSTAT.” Websites and Functionality. 2018. http://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-714389_QID_3257D732_UID_-3F171EB0&layout=TIME,C,X,0;SIZEN_R2,B,Y,0;GEO,B,Y,1;INDIC_IS,B,Z,0;UNIT,B,Z,1;INDICATORS,C,Z,2;&zSelection=DS-714389INDICATORS,OBS_FLAG;DS-714389UNIT,PC_ENT;DS-7143.

Eurostat, and OECD. 2005. *Oslo Manual: Guidlines for Collecting and Interpreting Innovation Data. Communities*. Vol. Third edit. OECD. https://doi.org/10.1787/9789264013100-en.

Fischer, Manfred M., and Arthur Getis. 2010. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Heidelberg, Berlin: Springer. https://doi.org/10.1017/CBO9781107415324.004.

Getis, Arthur. 2009. “Spatial Weights Matrices.” *Geographical Analysis* 41 (4): 404–10.

Gök, Abdullah, Alec Waterworth, and Philip Shapira. 2015. “Use of Web Mining in Studying Innovation.” *Scientometrics* 102 (1): 653–71. https://doi.org/10.1007/s11192-014-1434-0.

Grentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. “Text as Data.” 23276. NBER Working Paper Series. Cambridge, Massachusetts.

Jang, Seongsoo, Jinwon Kim, and Max von Zedtwitz. 2017. “The Importance of Spatial Agglomeration in Product Innovation: A Microgeography Perspective.” *Journal of Business Research* 78 (June). Elsevier: 143–54. https://doi.org/10.1016/j.jbusres.2017.05.017.

Katz, J Sylvan, and Viv Cothey. 2006. “Web Indicators for Compley Innovation Systems.” *Research Evaluation* 45 (5): 893–909. https://doi.org/10.1016/j.respol.2006.03.007.

Kerr, William R, Gilles Duranton, Ed Glaeser, and Vernon Henderson. 2014. “Agglomerative Forces and Cluster Shapes.” *Review of Economics and Statistics* 96 (3).

Kim, Jinhyung, Myunggwon Hwang, Do-Heon Jeong, and Hanmin Jung. 2012. “Technology Trends Analysis and Forecasting Application Based on Decision Tree and Statistical Feature Analysis.” *Expert Systems with Applications* 39 (16): 12618–25. https://doi.org/https://doi.org/10.1016/j.eswa.2012.05.021.

Kinne, Jan. 2018. “ARGUS - An Automated Robot for Generic Universal Scraping.” Mannheim: Centre for European Economic Research. https://github.com/datawizard1337/ARGUS.

Kleinknecht, Alfred, Kees Van Montfort, and Erik Brouwer. 2002. “The Non-Trivial Choice between Innovation Indicators.” *Economics of Innovation and New Technology* 11 (2): 109–21. https://doi.org/10.1080/10438590210899.

Kleinknecht, Alfred, and Jeroen O N Reijnen. 1993. “Towards Literature-Based Innovation Output Indicators.” *Structural Change and Economic Dynamics* 4 (1): 199–207. https://doi.org/10.1016/0954-349X(93)90012-9.

Krzywinski, Martin, and Naomi Altman. 2013. “Points of Significance: Significance, P Values and T-Tests.” *Nature Methods* 10 (11). Nature Publishing Group: 1041–42. https://doi.org/10.1038/nmeth.2698.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” https://doi.org/10.1162/153244303322533223.

Mikolov, Tomas, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. 2011. “Strategies for Training Large Scale Neural Network Language Models.” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.1109/ASRU.2011.6163930.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. “Linguistic Regularities in Continuous Space Word Representations.” In *Proceedings of NAACL-HLT*, 746–51. https://doi.org/10.3109/10826089109058901.

Miner, Gary, John Elder, Andrew Fast, Thomas Hill, Robert Nisbet, and Dursun Delen. 2012. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Cambridge, Massachusetts: Academic Press.

Nagaoka, Sadao, Kazuyuki Motohashi, and Akira Goto. 2010. “Patent Statistics as an Innovation Indicator.” In *Handbook of Economics of Innovation*, edited by Bronwyn H. Hall and Nathan Rosenberg, Vol. 2, 1083–1127.

Nathan, Max, and Anna Rosso. 2017. “Innovative Events.” 429. Centro Studi Luca d’Agliano Development Studies Working Paper. https://ssrn.com/abstract=3085935.

OECD. 2009. *OECD Patent Statistics Manual*. Paris: OECD. https://doi.org/10.1787/9789264056442-en.

———. 2015. “The Future of Productivity.” Paris. https://doi.org/10.1787/9789264248533-en.

———. 2017. “Broadband Portal.” Paris. www.oecd.org/sti/broadband/oecdbroadbandportal.htm.

Rammer, C, B Aschhoff, T Doherr, B. Peters, and T. Schmidt. 2017. “Innovationsverhalten Der Deutschen Wirtschaft.” *Indikatorenbericht Zur Innovationserhebung 2016*. Mannheim. http://ftp.zew.de/pub/zew-docs/mip/16/mip_2016.pdf.

Rammer, Christian, Jan Kinne, and Knut Blind. 2016. “Microgeography of Innovation in the City: Location Patterns of Innovative Firms in Berlin.” 16–080. ZEW Discussion Paper. Mannheim.

Raymond, Kosala, and Hendrik Blockeel. 2000. “Web Data Mining Research: A Survey.” *SIGKDD Explorations* 2 (1): 1–10. https://doi.org/10.1109/ICCIC.2010.5705856.

Scott, John, and Peter J. Carrington. 2011. *The SAGE Handbook of Social Network Analysis*. SAGE.

Scrapy Community. 2008. “Scrapy.” Scrapinghub Ltd. https://github.com/scrapy/scrapy.

Shepherd, William G., and Joanna Mehlhop Shepherd. 2003. *The Economics of Industrial Organization*. Long Grove, IL: Waveland Press Inc.

Squicciarini, Mariagrazia, and Chiara Criscuolo. 2013. “Measuring Patent Quality.” 2013/03. OECD Science, Technologya Dn Industry Working Papers. Paris. https://doi.org/http://dx.doi.org/10.1787/5k4522wkw1r8-en.

Steiger, Enrico, Bernd Resch, and Alexander Zipf. 2016. “Exploration of Spatiotemporal and Semantic Clusters of Twitter Data Using Unsupervised Neural Networks.” *International Journal of Geographic Information Science* 30 (9): 1694–1716.

Youtie, Jan, Diana Hicks, Philip Shapira, and Travis Horsley. 2012. “Pathways from Discovery to Commercialisation: Using Web Sources to Track Small and Medium-Sized Enterprise Strategies in Emerging Nanotechnologies.” *Technology Analysis and Strategic Management* 24 (10): 981–95. https://doi.org/10.1080/09537325.2012.724163.

Appendix

Table A1. Sectors' NACE code ranges.

NACE code range	Sector label	Level 1 codes
0-4999	Agriculture	A
5000-9999	Mining	B
10000-18999	Consumer goods	C
19000-20999	Petrochemistry	C
21000-21999	Pharmaceuticals	C
22000-24999	Materials	C
25000-25999	Metal products	C
26000-27999	Electronic products	C
28000-30999	Mechanical engineering	C
31000-34999	Other products	C
35000-40999	Public utility	D, E
41000-44999	Construction	F
45000-46999	Wholesale	G
47000-48999	Retail	G
49000-54999	Transport	H
55000-57999	Food services	I
58000-63999	ICT services	J
64000-68999	Financial services	K
69000-76999	Advanced services	M
77000-83999	Other personal services	M
84000-85999	Public services	O,P
86000-89999	Health/social services	Q
90000-99999	Other services	R

Table A2. OLS regression results: Number of webpages on firm's website.

Variable	Coefficient	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	-7.024	34.18
50-75%	-0.74	35.97
10-50%	5.67	36.35
0-10%	-64.75	93.85
Employees		
1-5	-684.28***	141.95
6-25	-632.31***	142.03
26-250	-467.18***	144.11
Age		
0-1	-173.20	99.27
2-5	-119.22	90.82
6-25	-148.30	87.75
26-100	-170.84	87.86
Sector		
Agriculture	-89.01	54.77
Mining	-52.25	81.12
Consumer goods	17.75	70.96
Petrochemistry	-116.27	61.11
Pharmaceuticals	-517.50**	191.47
Materials	70.27	101.29
Metal products	-99.09	59.21
Electronic products	178.60	111.76
Other products	4.55	73.68
Public utility	22.92	92.24
Construction	-61.70	54.02
Wholesale	139.49**	60.51
Retail	280.69***	67.66
Transport	-80.29	60.30
Food services	-26.76	59.25
ICT services	151.33**	70.75
Financial services	236.95***	71.76
Advanced services	32.50	56.68
Other personal services	25.59	62.51
Public services	166.28	91.29
Health/social services	-79.94	58.88
Other services	220.66***	68.35
Constant		
Constant	926.81***	178.63

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old.

*p≤0.05, **p≤0.01, ***p≤0.001

Table A3. OLS regression results: Mean number of characters per webpage.

Variable	Coefficient	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	-217.92	350.10
50-75%	522.53	1143.60
10-50%	-420.99	430.69
0-10%	134.47	517.62
Population density		
1,000 people/km²	328.08*	147.70
Employees		
1-5	-2010.08	2389.08
6-25	-2011.86	2394.10
26-250	-1748.15	2424.49
Age		
0-1	-47.64	796.93
2-5	125.15	709.39
6-25	176.11	606.41
26-100	-507.59	580.17
Sector		
Agriculture	-606.75	628.27
Mining	-1018.10	781.40
Consumer goods	-406.50	545.19
Petrochemistry	-932.72	726.01
Pharmaceuticals	-1320.89	1699.43
Materials	210.79	638.99
Metal products	-686.26	566.13
Electronic products	-684.93	618.31
Other products	-455.22	816.33
Public utility	1348.00	1092.19
Construction	-563.74	501.87
Wholesale	-86.04	552.02
Retail	868.69	741.74
Transport	255.92	714.85
Food services	-73.22	655.36
ICT services	3095.48	2756.91
Financial services	-152.54	547.71
Advanced services	455.85	536.17
Other personal services	643.74	698.96
Public services	-611.05	631.97
Health/social services	440.25	695.13
Other services	-296.41	512.15
Constant		
Constant	4984.50*	2435.30

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old.

*p≤0.05, **p≤0.01, ***p≤0.001; n=3,955

Table A4. OLS regression results: Share (0.0 to 1.0) of German language on firm's website.

Variable	Coefficient	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	-0.05	0.03
50-75%	-0.05	0.04
10-50%	-0.08*	0.04
0-10%	-0.01	0.07
Population density		
in 1,000 people/km²	-0.02	0.01
Employees		
1-5	0.03	0.07
6-25	-0.01	0.07
26-250	-0.02	0.07
Age		
0-1	-0.03	0.10
2-5	-0.14	0.08
6-25	-0.04	0.07
26-100	-0.01	0.06
Sector		
Agriculture	0.36*	0.15
Mining	0.43**	0.14
Consumer goods	0.27	0.17
Petrochemistry	0.21	0.18
Pharmaceuticals	-0.59***	0.14
Materials	0.29	0.18
Metal products	0.40**	0.14
Electronic products	0.14	0.22
Other products	0.30	0.19
Public utility	0.16	0.17
Construction	0.37**	0.14
Wholesale	0.36*	0.14
Retail	0.37	0.14
Transport	0.29*	0.15
Food services	0.27	0.15
ICT services	0.35*	0.15
Financial services	0.18	0.17
Advanced services	0.31*	0.15
Other personal services	0.37**	0.14
Public services	0.33*	0.17
Health/social services	0.41**	0.14
Other services	0.31*	0.15
Constant		
Constant	0.68***	0.17

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old.

*p≤0.05, **p≤0.01, ***p≤0.001; n=432

Table A5. OLS regression results: Mean number of hyperlinks per webpage.

Variable	Coefficient	Robust Std. Error
Broadband (≥50Mbits) availability		
76-95%	3.00	3.65
50-75%	10.22*	4.66
10-50%	7.53	4.59
0-10%	2.86	8.39
Population density		
in 1,000 people/km²	1.96	1.08
Employees		
1-5	-5.58	10.78
6-25	-3.37	10.68
26-250	-7.77	10.74
Age		
0-1	3.84	6.32
2-5	5.01	6.56
6-25	9.07	5.75
26-100	5.14	5.78
Sector		
Agriculture	-11.34	14.54
Mining	-18.06	14.46
Consumer goods	-10.16	13.79
Petrochemistry	-19.40	14.09
Pharmaceuticals	-14.08	14.34
Materials	-17.69	13.26
Metal products	-13.73	13.94
Electronic products	-21.28	13.22
Other products	-19.17	13.38
Public utility	-13.48	13.34
Construction	-4.41	14.34
Wholesale	-7.12	13.73
Retail	-7.35	15.05
Transport	-8.20	13.52
Food services	-4.17	14.16
ICT services	-9.01	13.92
Financial services	-3.82	13.99
Advanced services	0.71	16.00
Other personal services	2.02	16.52
Public services	-13.53	13.36
Health/social services	-4.71	13.94
Other services	-11.34	14.54
Constant		
Constant	19.85	17.80

Baseline firm: Mechanical engineering firm in region with >95% broadband availability, >250 employees, >100 years old.

*p≤0.05, **p≤0.01, ***p≤0.001; n=4,121

ARGUS Documentation

The free-to-use ARGUS web scraper and the full documentation can be found here:
https://github.com/datawizard1337/ARGUS

ARGUS

ARGUS is an easy-to-use web mining tool. The program is based on the Scrapy Python framework and is able to crawl a broad range of different websites. On these websites, ARGUS is able to perform tasks like scraping texts or collecting hyperlinks between websites.

Getting Started

These instructions will get you a copy of ARGUS up and running on your local machine.

Follow these 10 easy steps, which are described in more detail below, to carry out a successful ARGUS scraping run:

1. Install Python 3.6
2. Install additional Python packages.
3. Install cURL and add a cURL environment variable to your system.
4. Download and extract the ARGUS files.
5. Prepare the settings.txt and your list of website URLs.
6. Run the scrapyd server by double-clicking on "start_server.bat"
7. Start your scraping run by double-clicking on "start_scraping.bat"
8. Check your scraping run using the web interface.
9. Wait until all jobs have finished.
10. Run the "postprocessing.bat" and check out the results which were saved to the same directory your initial list of website URLs are located in.

System requirements

ARGUS works with Python 3.6, is based on the Scrapy framework and has the following Python package dependencies:

- Scrapy 1.5.0
- scrapyd 1.2.0
- scrapyd-client 1.1.0
- scrapy-fake-useragent 1.1.0
- tldextract 2.2.0

- pandas 0.22.0

Additionally, you need cURL to communicate with the ARGUS user interface. An executable Windows 64bit version of cURL can be downloaded here, for example. After downloading and extracting, you need to add a cURL environment variable to your system. See this Stackoverlow thread if you do not know how to do that.

Installation

If you are not using Python yet, the easiest way to install Python 3.6 and most of its crucial packages is to use the Anaconda Distribution. After installing Anaconda, you can use pip to install the packages above by typing "pip install package_name" (e.g., "pip install scrapy") into your system command prompt.

Using ARGUS

If you are interested in how ARGUS processes websites, read the following description of its workflow. Otherwise, you may continue with the next section.

An ARGUS crawl is based on a list of firm website addresses (URL) provided by the user and proceeds as follows:

1. The first webpage (a website's main page) is requested using the first address in the given URL list.
2. A collector item is instantiated, which is used to collect the website's content, meta-data (e.g. timestamps, number of scraped URLs etc.), and a so-called URL stack.
3. The main page is processed:
    - Content from the main page is extracted and stored in the collector item.
    - URLs which refer to subpages of the same website (i.e. domain) are extracted and stored in the collector item's URL stack.
4. The algorithm continues to request subpages of the website using URLs from the URL stack. Hereby, it can use a simple heuristic which gives higher priority to short URLs and those URLs which refer to subpages in a predefined language.
    - Content and URLs are collected from the subpage and stored in the collector item.
    - The next URL in the URL stack is processed.
5. The algorithm stops to process a domain once all subpages have been processed or as soon as a predefined number of subpages per domain have been processed.
6. The collected content is processed and exported to an output file.

7. The next website is processed by requesting the next URL from the URL list provided by the user. The process described above is repeated until all firm website addresses from the user list have been processed.

Spider types

Currently, ARGUS comes with two types of spiders: textspiders and linkspiders.

- **textspider** - these spiders extract texts from websites you give to them.
- **linkspider** - these spiders extract hyperlinks between websites you give to them. They also collect hyperlinks to websites "out-of-sample", but not between out-of-sample websites and from out-of-sample websites to within-sample websites.

The settings file

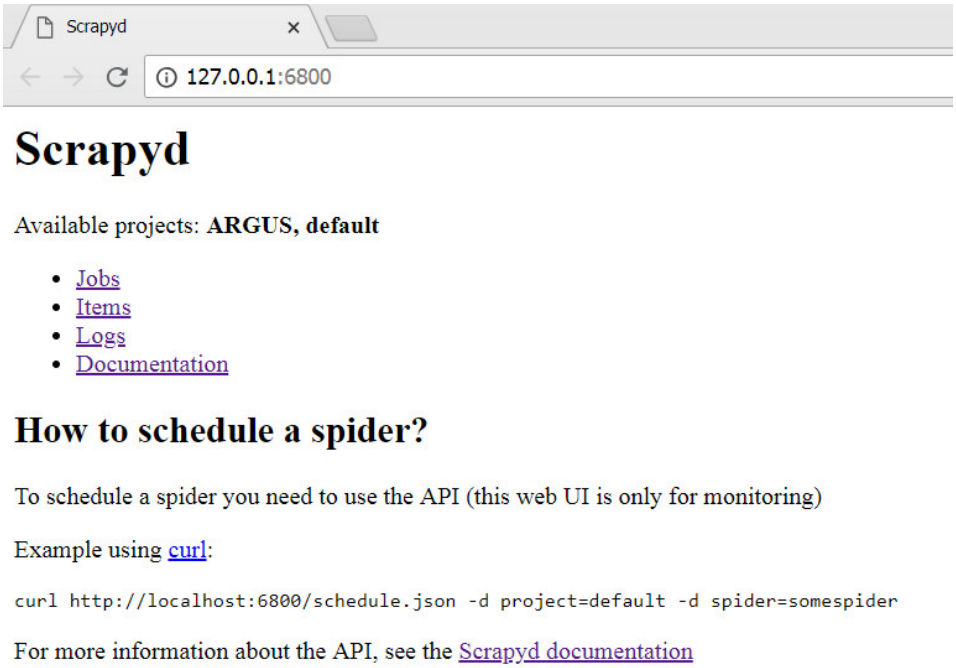
The first thing you have to do when performing an ARGUS crawl is to prepare the “settings.txt” which is located in the ARGUS root directory. In the settings file, the following parameters need to be set:

- [input-data]
    - **filepath** – the full path to your text file with website addresses. The file should be delimiter-separated and without BOM (byte order mark). An easy way to see whether your text file uses BOM is to use Notepad++ and check the “Encoding” in the top panel. The URLs need to be in the format “www.example.com”. The directory of your URL list will also be used to output the scraped data. An example website address can be found in /misc:
    - **delimiter** – the type of delimiter your text file uses. It is recommended to use tab-delimited text files: \t
    - **encoding** – the encoding of your text file. It is recommended to use text files in: utf-8
    - **ID** – the field name of your unique website identifier in your website address file.
    - **url** – the field name of your web addresses in your website address file.
- [system]
    - **n_cores** – the number of processor cores you want to dedicate to the ARGUS scraping process. It is recommended to use the total number of cores in your system -1 (i.e. if you have a quad-core processor with 4 cores, you should choose “n_cores = 3”).

- [spider-settings]
    - **spider** - select either *text* or *link* to use textspiders or linkspiders to process your websites.
    - **limit** – the maximum number of subpages (incl. the main/starting page) that will be scraped per domain. Set this to 0 if you want to scrape entire websites (caution is advised as there are websites with tens of thousands of subpages).
    - **prefer_short_urls** – whether you want ARGUS to preferentially download the shortest hyperlinks it finds on a website first. ARGUS usually starts at the website's main page where it collects all hyperlinks directing to the website's subpages. After processing the website's main page, ARGUS follows the hyperlinks it finds there and does the same to website's subpages until it reaches the set **limit**. If **prefer_short_urls** is set to "on", ARGUS will visit those subpages with the shortest URLs first. The reasoning behind this is that one can assume that the most general (and arguably most important) information is located at the website's top level webpages (e.g., www.example.com/products). If you want to turn this simple selection heuristic off, choose: off
    - **language** – the language that will be preferred when selecting the next subpage URL (analogous to **prefer_short_urls**). Note that this simple heuristic just checks the URL for certain ISO language codes. You need to insert the ISO language name as you find it in the "ISO_language_code.txt" in the ARGUS\misc sub directory. So if you wanted to prioritize German language URLs, you would enter: German. If you do not want to use this heuristic, just enter: None.
    - **log** – the amount of information that is stored in log files. The available options are DEBUG, INFO, WARNING, ERROR, and CRITITCAL. For larger scraping runs, the log level should be set to: INFO.

Starting a scraping run

Before starting your scraping run, a scrapyd server, which handles your scraping jobs, needs to be started. This can be done by running the "start_server.bat", which opens a separate window that should not be closed for the entirety of the upcoming scraping run. After the server has started, the scraping process can be launched by executing "start_scraping.bat". This little program will split your list of URLs into handy chunks and starts a separate job for each chunk to speed up the scraping process. The splitting and job scheduling may take a short while. After all jobs have been scheduled, the scrapyd web interface will open up in your default web browser (you can also get there by typing "http://127.0.0.1:6800/" into your web browser).



You can safely ignore the lower part about how to schedule a spider, because ARGUS does that for you. To see the jobs which have been scheduled, click the "Jobs" link. There you will find an overview about the pending, running, and finished jobs. You can also see the time a job was started, its current runtime, and the time it was finished. By clicking on a job's log link, you can have a look at its log file. The number of running jobs should be equal to the **n_cores** parameter you set in the "settings.txt".

Stopping jobs

Sometimes certain jobs stop working or never finish, so you may want to stop and restart them. This can be done by running the "kill_single_job.bat". You will be asked for the id of the job you want to cancel. The id is a long hash number which can be found in the "Job" column in the "Jobs" web interface section.

Project	Spider	Job	PID	Start	Runtime	Finish	Log	Items
Pending								
Running								
ARGUS	textspider	24ca036873b311e8b65d5459596fa410	6040	2018-06-19 13:23:09	0:00:02		Log	Items
ARGUS	textspider	2508e12873b311e88da95459596fa410	8356	2018-06-19 13:23:09	0:00:02		Log	Items
Finished								

You can stop all processes at once by running the “kill_all_jobs.bat”. This little program will tell the scrapyd server to stop all running and scheduled processes. You will be asked whether you want to delete the data already scraped. If you decide against deleting the scraped data, you may want to run the “postprocessing.bat” as described below.

Postprocessing

When all jobs are finished, you may close the scrapyd server window to stop the server. Finally, you need to run “postprocessing.bat” which cleans up and writes your scraped data to the directory of your input data.

Output data

The output file can be found in the same directory your original website address file is located (**filepath** parameter in the settings file).

Textspider output

One row equals one webpage and n (n ≤ **limit**) webpages equal one website (identified by its ID).

	ID	dl_rank	dl_slot	error	redirect	start_page	text	timestamp	url
1471	313455	0	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/
1472	313455	1	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/
1473	313455	2	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/rss/
1474	313455	3	zew.de	None	False	https://www.zew.de/	Das Zentrum für Europäische Wirtschaftsforschu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/team/
1475	313455	4	zew.de	None	False	https://www.zew.de/	Unser informiert Sie direkt über Neuigkeiten a...	Wed May 23 13:06:30 2018	https://www.zew.de/de/presse/
1476	313455	5	zew.de	None	False	https://www.zew.de/	Tel: 0621 1235-132 Fax: 0621 1235-255 E-Mail: ...	Wed May 23 13:06:30 2018	https://www.zew.de/de/kontakt/
1477	313455	6	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/team/ybr
1478	313455	7	zew.de	None	False	https://www.zew.de/	Das ZEW ist ein gemeinnütziges wirtschaftswiss...	Wed May 23 13:06:30 2018	https://www.zew.de/de/das-zew/
1479	313455	8	zew.de	None	False	https://www.zew.de/	Wenn Sie den Hauptbahnhof verlassen haben, übe...	Wed May 23 13:06:30 2018	https://www.zew.de/de/anfahrt/

- **ID** – the ID of the website as given in [input-data] section of the settings file.
- **dl_rank** – the chronological order the webpage was downloaded. The main page of a website (i.e. the URL in your website address file) has rank 0, the first subpage processed after the main page has rank 1, and so on.
- **dl_slot** – the domain name of the website as found in the user given website address list.
- **error** – not “None” if there was an error requesting the website’s main page. Can be an HTML error (e.g., “404”), DNS lookup error, or a timeout.

- **redirect** – is “True” if there was a redirect to another domain when requesting the first webpage from a website. This may indicate that ARGUS scraped a different website than intended. However, it may also be a less severe redirect like “www.example.de” to “www.example.com”. It is your responsibility to deal with redirects.
- **start_page** – gives you the first webpage that was scraped from this website. Usually, this should be the URL given in your website address file.
- **text** – the text that was downloaded from the webpage.
- **timestamp** – the exact time when the webpage was downloaded.
- **url** – the URL of the webpage.

Linkspider output

One row equals one website (identified by its ID). All hyperlinks found on n (n ≤ **limit**) webpages are aggregated to the website (domain) level and duplicates are removed.

	ID	dl_rank	dl_slot	error	redirect	start_page	text	timestamp	url
1471	313455	0	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/
1472	313455	1	zew.de	None	False	https://www.zew.de/	Wie beeinflussen aktuelle politische Entwicklu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/
1473	313455	2	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/rss/
1474	313455	3	zew.de	None	False	https://www.zew.de/	Das Zentrum für Europäische Wirtschaftsforschu...	Wed May 23 13:06:30 2018	https://www.zew.de/de/team/
1475	313455	4	zew.de	None	False	https://www.zew.de/	Unser informiert Sie direkt über Neuigkeiten a...	Wed May 23 13:06:30 2018	https://www.zew.de/de/presse/
1476	313455	5	zew.de	None	False	https://www.zew.de/	Tel: 0621 1235-132 Fax: 0621 1235-255 E-Mail: ...	Wed May 23 13:06:30 2018	https://www.zew.de/de/kontakt/
1477	313455	6	zew.de	None	False	https://www.zew.de/	NaN	Wed May 23 13:06:30 2018	http://www.zew.de/de/team/ybr
1478	313455	7	zew.de	None	False	https://www.zew.de/	Das ZEW ist ein gemeinnütziges wirtschaftswiss...	Wed May 23 13:06:30 2018	https://www.zew.de/de/das-zew/
1479	313455	8	zew.de	None	False	https://www.zew.de/	Wenn Sie den Hauptbahnhof verlassen haben, übe...	Wed May 23 13:06:30 2018	https://www.zew.de/de/anfahrt/

- **ID** – the ID of the website as given in [input-data] section of the settings file.
- **alias** – if there was an initial redirect (e.g. from www.example.de to www.example.com), the domain the spider got redirected to ("example.com" in the example) becomes the website’s alias.
- **dl_slot** – the domain name of the website as found in the website address list provided by the user.
- **error** – not “None” if there was an error requesting the website’s main page. Can be an HTML error (e.g., “404”), DNS lookup error, or a timeout.
- **links_internal** – the domains of "within-sample" websites found on the focal website. The first element is the focal website itself (this format makes it easier to import the data as an "adjacency list" into analysis software). Field is empty if no hyperlinks to within-sample websites were found.
- **links_external** – the domains of "within-sample" and "out-of-sample" websites found on the focal website. The first element is the focal website itself (this format makes it easiert

to import the data as an "adjacency list" into analysis software). Field is empty if no hyper-links were found.

- **redirect** – is “True” if there was a redirect to another domain when requesting the first webpage from a website. This may indicate that ARGUS scraped a different website to the one intended. However, it may also be a less severe redirect like “www.example.de” to “www.example.com”. It is your responsibility to deal with redirects.
- **timestamp** – the exact time when the webpage was downloaded.
- **url** – the URL of the webpage.

Why ARGUS?

ARGUS stands for "**A**utomated **R**obot for **G**eneric **U**niversal **S**craping".