// NO.18-033 | 12/2019

DISCUSSION PAPER

// JAN KINNE AND
JANNA AXENBECK

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

Leibniz Association

ZEW

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

Jan Kinne[1,2,3]* and Janna Axenbeck[4,5]

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

[2] Z_GIS - Department of Geoinformatics, University of Salzburg

[3] Center for Geographic Analysis, Harvard University

[4] Department of Digital Economy, ZEW - Centre for European Economic Research

[5] Department of Economics of Digitalisation, Justus-Liebig-University Giessen

* Corresponding author. Mail: jan.kinne@zew.de

Abstract: Existing approaches to model innovation ecosystems have been mostly restricted to qualitative and small-scale levels or, when relying on traditional innovation indicators such as patents and questionnaire-based survey, suffered from a lack of timeliness, granularity, and coverage. Websites of firms (as well as research institutes and universities, which are not part of this study) are a particularly interesting data source for innovation research, as they are used for publishing information about potentially innovative products, services, and co-operation with other firms. Analyzing the textual and relational content on these websites and extracting innovation-related information from them has the potential to provide researchers and policy-makers with a cost-effective way to survey millions of businesses via their web-sites, gain insights into their innovation activity, their cooperation, and applied technologies. For this purpose, we propose a web mining framework for consistent and reproducible map-ping of innovation ecosystems. In a large-scale pilot study we use a database with 2.4 million German firms to test our framework and explore firm websites as a data source. Thereby we put particular emphasis on the investigation of a potential bias when surveying innovation systems through firm website if only certain firm types can be surveyed using our proposed approach. We find that the availability of a websites and the characteristics of the website (number of subpages and hyperlinks, text volume, language used) differs according to firm size, age, location, and sector. We also find that patenting firms will be overrepresented in web mining studies. Web mining as a survey method also has to cope with extremely large and hyper-connected outlier websites and the fact that low broadband availability appears to prevent firms from operating their own website and thus excludes them from web mining analysis. Finally, we outline several approaches how to transfer firm website content into valuable innovation indicators.

Keywords: Web Mining; Web Scraping; Innovation

JEL Classification: O30, C81, C88

Past version: August 2018 **This version:** December 2019

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. We would also like to thank Sebastian Schmidt for his contribution to the development of ARGUS.

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 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 innovation ecosystems. 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 innovation 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 innovation ecosystems (Nagaoka, Motohashi, & Goto, 2010; OECD, 2009; Squicciarini & Criscuolo, 2013).

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 (OECD & Eurostat, 2018) 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 their R&D expenditures. Furthermore, they characterize an innovation by its degree of novelty (new to the firm, the market, the industry or the world) and the type of innovation (product, process, marketing, and organizational innovations). However, such indicators suffer from some major drawbacks. The German 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 extrapolation. Furthermore, rare but potentially important innovation activities happening in unobserved sectors or technological fields may not be covered in the data at all. This also affects the analysis of geospatial innovation processes, some of which happen to operate on a fine (micro-)geographical scale (Arzaghi & Henderson, 2008; Carlino & Kerr, 2015; Catalini, 2012; Jang, Kim, & von Zedtwitz, 2017; Kerr, Duranton, Glaeser, & Henderson, 2014). Consequently, established innovation indicators from questionnaire-based surveys lack sectoral, technological, and geographical 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, & Brouwer, 2002).

As an alternative to questionnaire-based surveys, innovation activity has been 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 & Pianta, 1996). Moreover, most patents are never used (Shepherd & Shepherd, 2003); thus, they serve rather as indicators of inventions than of innovations. Another drawback of patent-based indictors, especially if they take a more selective approach, is that the dataset suffers from insufficient timeliness (Squicciarini & 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 & Reijnen, 1993). In addition, Acs, Anselin, and Varga (2002) indicate that LBIOs under-represent innovations in smaller firms as their presence in the media is usually smaller.

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

- *Coverage*: They cover only a fraction of the overall firm population.
- *Granularity*: They suffer from insufficient sectoral, technological, and geographical 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.

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 (Askitas & Zimmermann, 2015; Raymond & Blockeel, 2000).

In economic research and ecosystem mapping, 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, & Shapira, 2015). Surveying firms using their websites instead of conducting interviews or questionnaires or using other traditional methods, offers some clear advantages (scale, cost, timeliness of the survey), but also comes with its own challenges (challenging data collection, data harmonization, and data analysis). However, no consistent approach for studying firm websites has been established yet. 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.

In this paper, we develop and present a coherent web mining framework that is based on *ARGUS* (***A**utomated **R**obot for **G**eneric **U**niversal **S**craping*), an easy and free-to-use web scraping tool which allows for large-scale data retrieval from websites without requiring the user to have expert knowledge of web scraping technology. We then apply ARGUS in a pilot study using the entire firm population of Germany. The aim of this pilot study is to investigate and quantitatively assess firm websites as a data source for web-based innovation indicators and innovation ecosystem mapping, as well as to derive best practice guidelines for researchers who use ARGUS for large-scale web surveys. The following two research questions guideline our pilot study:

- **Research Question 1** *URL Coverage*: What subpopulation of firms can be surveyed using web mining of firm websites and is a systematic bias in terms of firm characteristics (age, size, sector, location etc.) to be expected?
- **Research Question 2** *Website Characteristics*: How do firm websites differ in terms of their size and content and how does that interfere with web mining studies?

The remainder of this paper is organized as follows. First, we summarize the results of previous innovation research studies that used web mining. In the following Methods section, we present our web mining framework and the ARGUS web scraping tool. In section 4, we present our data. The results of our pilot study are presented in section 5 and are discussed in section 6. Section 7 concludes and outlines future research.

2. Previous Research

There are only a few existing studies analyzing the usability of web-based innovation indicators and web mining for innovation ecosystem modelling. 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 & Cothey, 2006, p. 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. Other authors used such an approach in combination with visual network-based methods to identify business deals, funding relations, and alliances (Basole, Huhtamäki, Still, & Russell, 2016; Basole et al., 2015; Rubens, Still, Huhtamaki, & Russell, 2011).

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 innovation ecosystems 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.

3. 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*.

3.1. A web mining framework for mapping innovation ecosystems

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 et al., 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 which can be used to generate innovation indicators.

In Figure 1, we outline such a methodology in the form of a general analysis framework for mapping innovation ecosystems and generating web-based firm-level innovation indicators. Similar to traditional innovation indicators, the base data is a firm 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 auxiliary databases containing established innovation indicators from questionnaire-based surveys, firm-level patenting data or literature data (LBIO), such that traditional innovation indicators are available for a subsample of the firms in the main dataset. 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 down-

loaded 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 new 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.

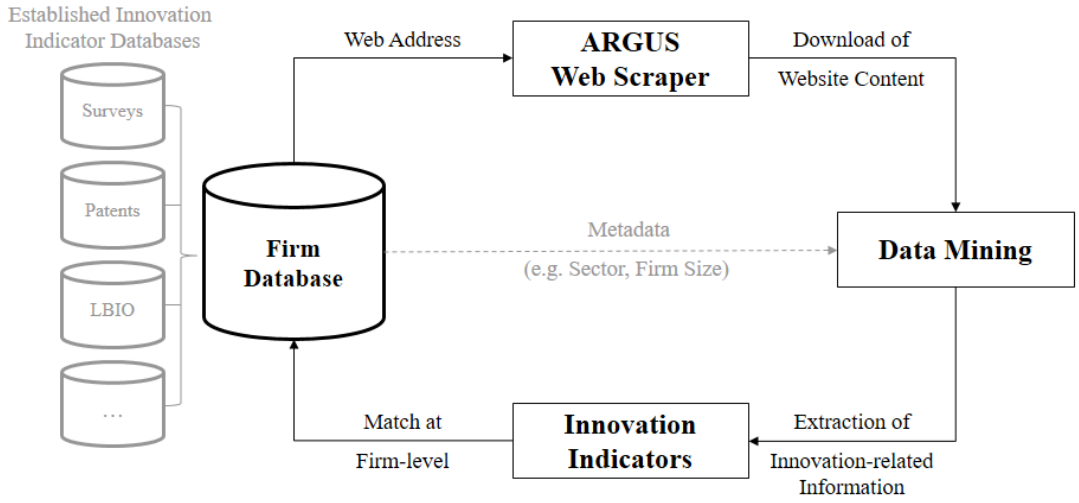


Figure 1. General analysis framework for mapping innovation ecosystems.

The proposed analysis framework allows for an automated, less costly mapping of entire firm populations that can be carried out faster and in shorter time intervals in comparison to traditional approaches. Also, this approach is easily expandable to map knowledge ecosystems (see e.g. Xu, Wu, Minshall, & Zhou, 2018) by scanning the websites of universities and research institutes. Furthermore, receiving firm information from websites does not require any effort on the part of the analyzed firms. As a result, web-based indicators created this way 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. Recent technological and methodological advances in analyzing unstructured data using machine learning (Grentzkow, Kelly, & Taddy, 2017; Mikolov, Deoras, Povey, Burget, & Cernocky, 2011; Steiger, Resch, & Zipf, 2016) may have that potential. 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 may be to extract interpretable and meaningful information on firms' innovation activities (see Conclusion and Future Research section).

3.2. ARGUS web scraper

ARGUS (**A**utomated **R**obot for **G**eneric **U**niversal **S**craping) is a web scraping software tool that was developed to meet the requirements that are determined by the web mining framework outlined in the previous section:

- **Adaptability**: The web scraper must be able to scrape a wide variety of web content from any website. At the same time, the web scraper's output must be in a structured and consistent format.
- **Scalability**: The web scraper must be able to scrape tens of millions of webpages from millions of firm websites in a reasonable time frame that allows for frequent iterations of the scraping process in order to build up a panel database of web data.
- **Easy-to-use**: The web scraper must be easy-to-use such that it can be used by researchers without profound knowledge in web scraping technology.
- **Free and Open Source**: In order to ensure a rapid dissemination as well as a sustainable further development of the web scraper, the program must be free-to-use and open source.

ARGUS is based on the Scrapy Python framework (Scrapy Community, 2008) and is available open source via Github (Kinne, 2018). The program features a graphical user interface (see Figure 2) that allows for a rather easy and command line free control.

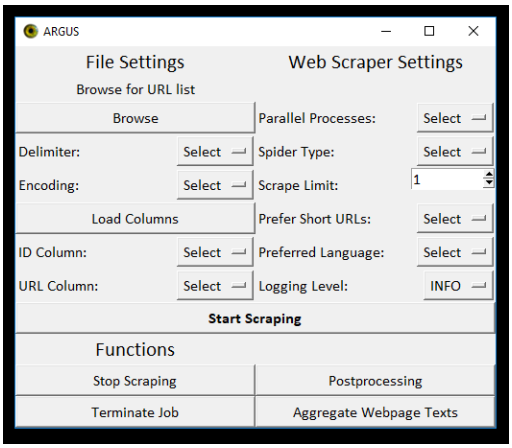


Figure 2. ARGUS graphical user interface.

4. Data

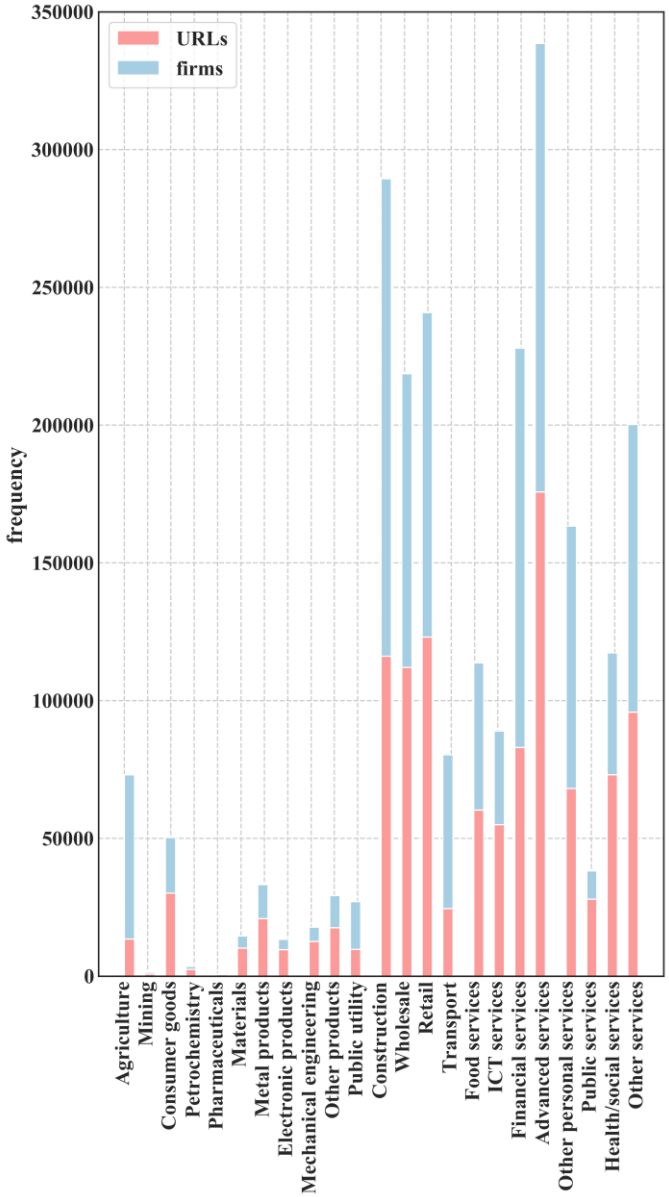
For the pilot study conducted in this paper, we use the *Mannheim Enterprise Panel* (MUP) as our base firm dataset. The MUP is a panel database that covers the total population of firms located in Germany. It contains about three million firm observations which are updated on a semi-annual basis. We restrict the dataset to firms that were definitely economically active in 2018 (2.52 million firms). The dataset also includes 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).

Patents are one of the most widely used and established innovation indicators (see e.g. Acs, Anselin, & Varga, 2002b; Archibugi & Pianta, 1996; Griliches, 1990; Nelson, 2009; OECD, 2009). We gathered patent data (patent stock end of 2017) from the European Patent Office and conducted a firm-patent match with our MUP firm database. Thereby, we restricted the patent dataset to patents that were filed after 2005 (10 years is the average lifetime of a patent in our database) to account for the decreasing economic and technological value of aging patents (Behrens, Hünermund, Leitner, Licht, & Peters, 2018).

5. Results

5.1. URL coverage

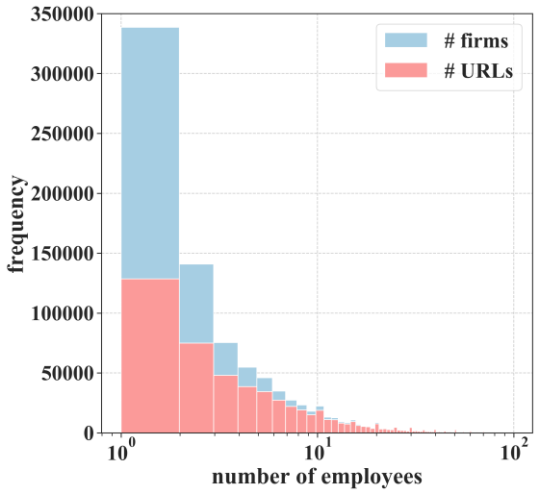
The overall URL coverage in our dataset is at 46% (1.15 million firms), but differs with firm size, sector, and location. Table 1 shows a breakdown of the firm population and URL coverage by sectors (a NACE code to sector mapping can be found in Table A1 in the appendix). Some sectors have a considerably higher URL coverage (≥70% coverage for materials, electronic products, mechanical engineering, and public services) than others (≤40% coverage for agriculture, public utility, construction, transport, financial services).



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 1. URL coverage by sector.

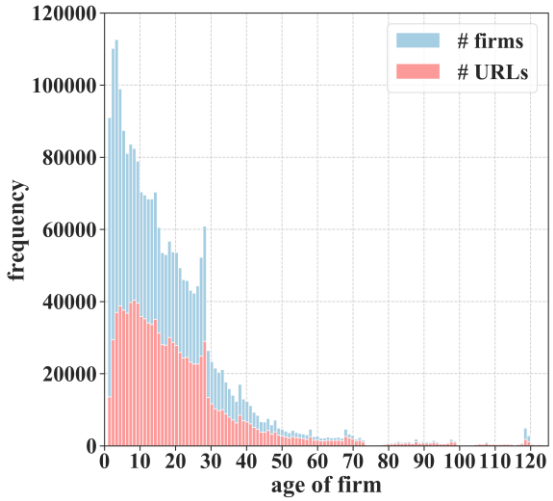
Table 2 shows firms' URL coverage by firm size groups (number of employees; variable available for 38% of firms). We can see that most firms are very small (micro-enterprises with less than 6 employees) and that coverage for this group is rather low (49%). For small firms (6-25 employees) coverage is decent (84%). Medium (26-250 employees) and large firms (>250 employees) are covered very well (94% and 97% respectively). 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 & 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).



# 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 2. URL coverage by firm size.

Table 3 shows firms' URL coverage by age (variable available for 91% of firms). Several historical events with an increased founding activity can be seen in the distribution (left panel): 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). A trend of increasing URL coverage with firm age is visible: 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 positively correlated (Spearman's rho of 0.37; p<0.001). A two-sample t-test indicated a highly significant difference between the age the overall firm population ($\bar{x}$=16.7) and the URL covered subpopulation ($\bar{x}$=21.2).



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**

Table 3. URL coverage by firm age.

Figure 4 maps the ratio of firms with an available URL to the overall local firm population by district. Low and high ratios do not seem to be randomly scattered, but instead low coverage can be primarily found in the East of Germany, while the Western part seems to be well covered. This impression of non-randomness is confirmed by a high and significant *Moran's I* (see e.g. Fischer & 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 and low 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 at the level of districts (Spearman rho of 0.5; $p<0.001$).

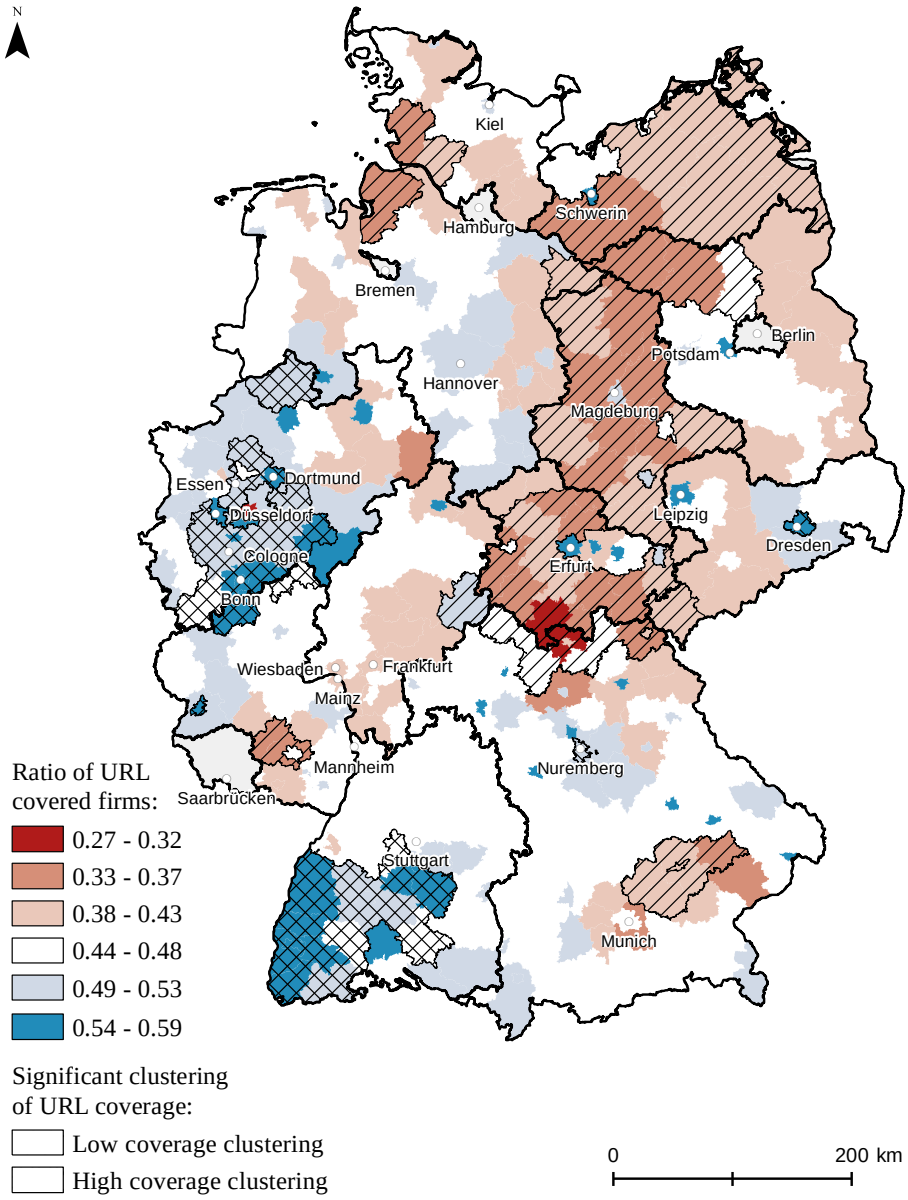


Figure 4. URL coverage by districts.

We investigate the relationships between the discussed firm characteristics and the availability of a URL in a probit regression analysis. The regression analysis results (as marginal effects) are shown in Table 4. *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, & 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.

Missing URLs in our data can result from either incomplete inquiry by our data provider or the fact that firms have actually no website. We investigate this issue by including two control variables in the regression analysis. Some legal forms do require a mandatory entry in official commercial registries – a procedure which makes surveying the firm a lot easier and, thus, likely 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 a possible bias in our data provider's search strategy too. 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.

The baseline firm in the regression 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 (VIF) is 9.36, which may indicate problematic multicollinearity in our model (the corresponding correlation Table A2 can be found in the appendix). While some authors emphasize a VIF of lower than 10 (Kutner, Nachtsheim, Neter, & Li, 2005), others suggest a significantly lower threshold of 3 (Tabachnick & Fidell, 2006).

Overall, the findings from the descriptive statistics are confirmed by the probit regression. Very young and very small firms do not have websites and the sector plays an important role. 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 running their own website. According to our estimated effects, 30,000 firms in Germany (extrapolated to the total firm population) do not have a own 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. Dependent variable: Available firm website URL (yes/no).

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

Overall, 17,294 firms (0.6% of all firms) in our MUP dataset are patent holders and 71.47% of them are covered by a URL. Such a high URL coverage of patent holder firms was to be expected, given that mainly larger firms from sectors with a high URL coverage hold patents. As a result, patent holder firms will be overrepresented in web mining studies (1.3% patent holders after scraping compared to 0.6% in our base dataset). Figure 5 shows a breakdown of the share of patent holder firms by sector. While there is no eye-catching difference in the sector-level URL coverage of patent holder firms, the figures does highlight a well-known shortcoming of patents as innovation indicators. While patents play a crucial role to protect intellectual property in some sectors like mechanical engineering and pharmaceuticals other sectors where many firms may be considered as innovative patents do not fulfil this role. In the ICT services sector, for example, only 0.8% if firms hold patents, which is attributable to the fact that software is not patentable in Germany.

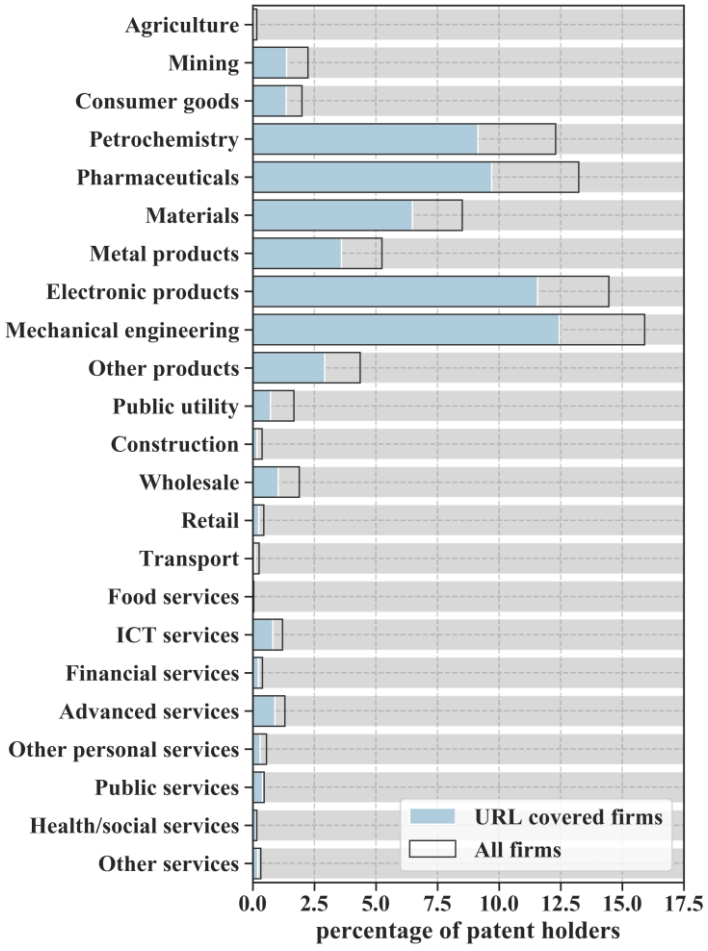


Figure 5. Share of patent holders by sector.

5.2. Website characteristics

For our further in-depth analysis of firm website characteristics, we randomly sampled 11,477 firms with a URL from our dataset and used ARGUS to scrape their websites. 84.2% of the websites could be scraped, while the he remaining 15.8% 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.

We then 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 redirects to the website of its 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 and excluded 9.5% of the URLs that were successfully crawled but were also tagged as redirected. T-tests showed no significant difference in firms' age and size between redirecting and non-redirecting URLs. In sum, 23.8% of firms had to be excluded from further analysis due to redirect or request errors, reducing our sample to 8,744 firms.

For the remaining firms, the mean number of webpages per website is 218.8 (SD 604.7) and the median is 15, resulting in a highly skewed distribution, as it can be seen in Figure 6. A considerable share (5.86%) of the websites reached the *Scrape Limit* (see Methods section) of 2,500 subpages which we set for this analysis. Differences between sectors are stark as seen in Figure 7, where the mean number of webpages (indicated as red dots) vary considerably between sectors. Some of this variation is due to the positive correlation (Spearman's rho of 0.19; $p<0.001$) between firm size (which also varies systematically with the sector) and the number of webpages on a firm's website.

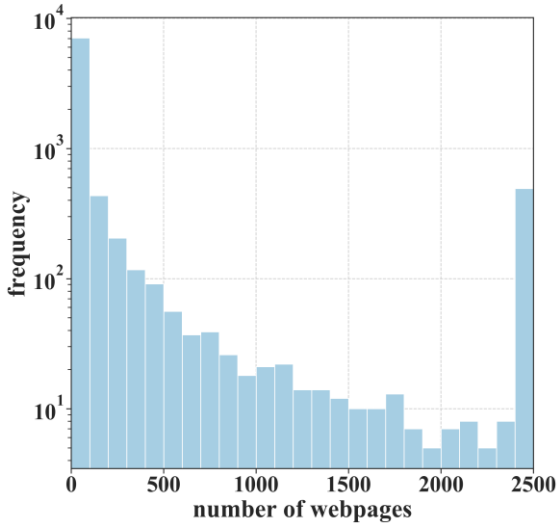


Figure 6. Number of webpages on firm website.

Figure 7. Number of webpages on firm website by sectors.

On average, a webpage we have downloaded has 3295.86 characters (SD=9960.43) and half of them have 1970 characters or less (which equals about two thirds of a standard page of text), resulting in a highly skewed distribution as it seen in Figure 8. We did not find any statistically significant relationship between the mean text length per webpage and any firm characteristic.

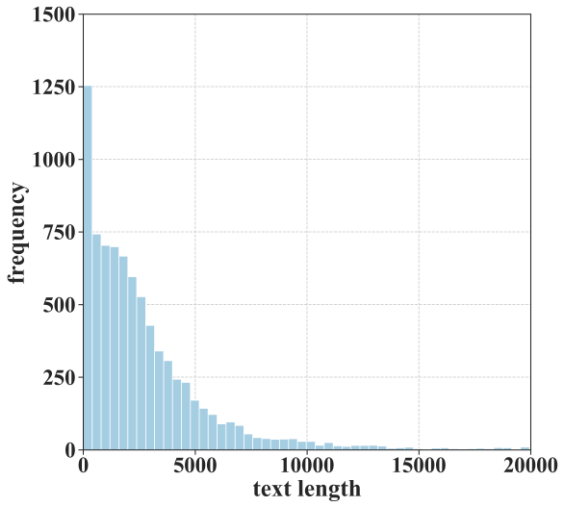


Figure 8. Mean text length per webpage.

We randomly sampled 911 websites and used Python's langdetect library (Danilak, 2015) to identify the languages used in each of their 193,504 sub-webpages. The algorithm was able to classify 91.9% of these webpages of which 88.2% were classified as being written in German. Most (60.8%) of the non-German language webpages were classified as written in English. Most of the firms have websites that are written almost completely in German (close to 100% of their webpages were classified as German), as it can be seen in Figure 9. Some firms only have non-German texts on their websites (share < 0.2; 4.5%). Figure 10 shows that the share of German language on a firm's website is related to the firm's sector (we do not show sectors with fewer than 10 observations). We do not find any other statistically significant relation to other firm characteristics.

It is important to keep in mind that sub-webpages were not selected uniformly or randomly from the firms' websites, as we used ARGUS' language selection heuristic set to German. Consequently, if a firm website was classified to be completely in German that does not automatically imply that the firm uses German exclusively on its website. Changing the preferred language from German 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 both German and English versions of their

website and ARGUS is indeed able to scrape a preferred language – a desirable feature as most natural language processing methods require text corpora in a single language.

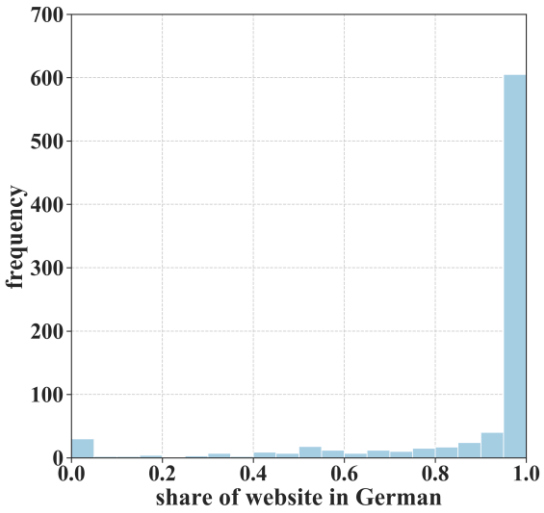


Figure 9. Share of website in German.

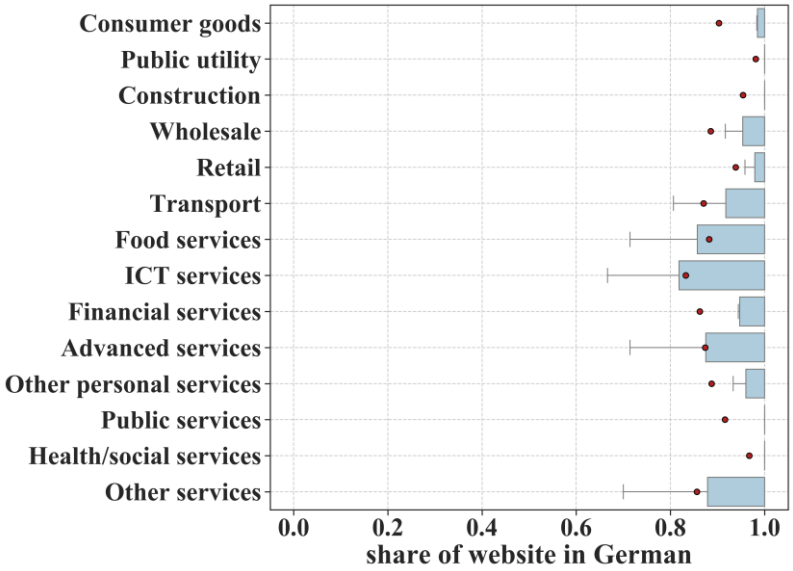


Figure 10. Share of website in German by sectors.

We also investigated the number of hyperlinks that connect a website to other websites in the World Wide Web by scraping our random sample of 11,477 firms using ARGUS' hyperlink scraping mode (*Scrape Limit* set to 100). We found that no website has less than 14 hyperlinks to other websites and some outlier websites have tens of thousands of such connections. 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. its number of sub-webpages). Looking at the mean number of hyperlinks per webpage, we see that, on average, a webpage contains 14.52 hyperlinks. The median number of hyperlinks per webpage is just 6, resulting in a highly skewed distribution as it can be seen in Figure 11. We did not find statistically significant relationships between the number of hyperlinks per webpage and any firm characteristics.

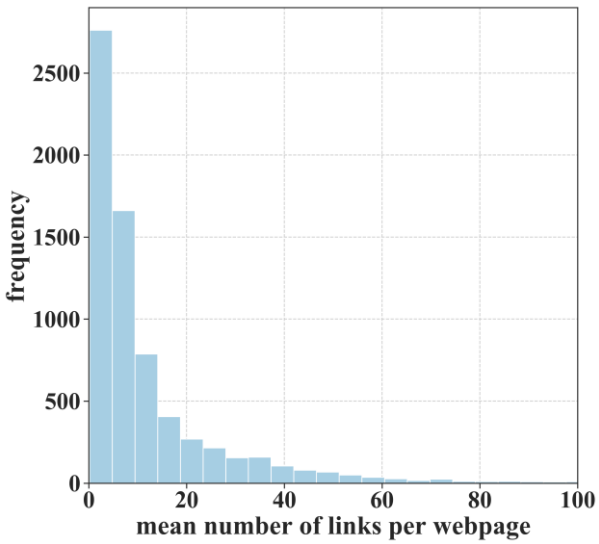


Figure 11. Mean number of hyperlinks per webpage.

6. Discussion

In the first part of our study, we investigated what firms in the total population of firms actually have their own websites (URL coverage) which would allow researchers to survey them in a web-based study. Thereby, we put particular emphasis on firm characteristics and their statistical relations to the URL coverage in the overall firm population. For this purpose, we also tried to untangle the cause of missing URLs in our firm 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 our data provider). Based on our case study results, regularities in URL coverage remain after controlling for a potential bias in the search strategy of our data provider. Researchers who conduct web mining to map innovation ecosystems as we proposed it in our framework will have difficulties observing very young and very small firms, especially those from certain sectors such as agriculture and those located in rural areas. In addition, low broadband availability seems to deter firms from setting up their own

website and therefore systematically excludes them from any web-based studies. If one assumes that low broadband availability is associated to a generally lower use of the Internet (both private and commercial) in a region, this may actually indicate that firms with local target markets that are located in an area with a low broadband availability have no incentive to set up their own website in order to communicate with their 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 web mining framework. This is especially true in urban areas. Given that the vast majority of innovative activity in Germany is conducted by the latter firm type (Rammer, Aschhoff, Doherr, Peters, & Schmidt, 2017), we can conclude that our web mining framework is suitable for analyzing the most important business-side parts of the German innovation ecosystem. This assumption is backed by our finding that patenting firms are overrepresented in web mining studies due to the higher URL coverage in patent-intensive firm subgroups.

We identified URL redirects as a potential issue when conducting web mining studies because outdated URLs can result in potentially harmful redirects. If conducting a large-scale web study based on a huge firm datasets, it is usually not possible to make sure that the available firm website addresses are all up-to-date. To minimize the share of erroneous scraped we content, we therefore recommend excluding firms such URL redirects. Given that less than 10% of successful URL requests were redirected and we did not find any systematic firm age or size bias, such an exclusion seems reasonable.

Our results showed that firm website size is highly correlated to firm size (number of employees) and sectors. Large firms have both more webpages on their websites and more text on each of these webpages. In general, we find that outliers play an important role when conducting web mining studies. Some websites are extremely large in terms of the number of webpages and the amounts of text provided on them. 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 completely scrape two thirds of all firm websites, it is therefore sufficient to set the limit of downloaded webpages per website to 50. If this threshold is increased to 250, 90% of the websites can be scraped entirely. About 6% of firms can be seen as extreme outliers with 2,500 or more sub-webpages on their websites.

Based on these purely quantitative results, it is difficult to make any generally applicable best practice recommendation for an appropriate *Scrape Limit* for ARGUS. If researchers are interested in generating a more general textual description of the firms, they may select a

rather low Scrape limit of 15 and would still scrape half of all firm websites entirely. If they are interested in highly specific information, that may be located on lower levels of the website, the need to set a rather high scrape limit around 250. In this sense, our results should provide researchers with a sound reference point when conducting their own web mining studies.

Unsurprisingly, our results showed that 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. We were also able to show that the ARGUS simple language selection heuristic helps to restrict the downloaded texts downloaded to a certain language. 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 a considerable share of firms provide several versions of their website in different languages. The language selection heuristic of ARGUS 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 firm websites is, unsurprisingly, highly correlated to the number of webpages it has. The mean number of links per webpage, however, seems to be randomly distributed with no significant relationship to the firm 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. regional vs. transregional) compared to other firm types (e.g. non-innovative firms).

7. Conclusion and Future Research

7.1. Conlusion

In this paper, we proposed a web mining framework for the mapping of innovation ecosystems by generating innovation indicators from website contents. We argued that established innovation indicators have a number of shortcomings concerning their coverage, granularity, timeliness, and data collection costs and that web-based indicators have the potential to overcome some of these limitations. The proposed web mining framework is composed of four key parts: a firm database with firm-level metadata and the firms' web addresses, ARGUS web scraper which is used to download firm website content, a data mining part to extract innovation-related information from the downloaded web content, and the actual innovation indicators generated from the extracted information. In the remainder of the paper we conducted a large-scale pilot study to investigate firm websites as a potentially valuable data source for innovation ecosystem mapping. Two research questions were the guideline for this pilot study.

- **URL coverage:** URL coverage (the availability of a website for a firm) differs systematically with firm characteristics. Certain types of firms can, thus, not be surveyed using our proposed web mining framework. Especially very young and very small firms, as well as firms from certain sectors and regions exhibit a very low URL coverage. Furthermore, we find that low local broadband availability can prevent firms from setting up their own internet presence. On the other hand, we find that almost all medium to large sized firms from sectors such as mechanical engineering and ICT services have websites. We also found that URL coverage is especially high among patenting firms. Given that the vast majority of innovative activity in Germany is conducted by these firm types, we can conclude that our web mining framework is suitable for analyzing the most important parts of the firm innovation systems.
- **Website characteristics:** We concluded that web mining studies have to deal with outlier issues. About 6% of firm websites have a number of sub-webpages four or more standard deviations above the population mean. Concerning the number of hyperlinks and the text volume found on these websites, this issue is even more evident. Large firms do not only operate larger websites, they also provide disproportionally more hyperlinks and text on them. We also found that there are sectoral differences concerning the size of firm websites and the languages used on them. We were also able to show that the language selection

heuristic of ARGUS effectively restricts text downloads to a certain language, which allows users to leverage the fact that many firms provide several versions of theirs websites in different languages. An important feature given that most natural language processing methods require texts in a single language.

7.1. Future Research

In future research, the analysis of the downloaded web data and the inclusion of other subsystems of the innovation ecosystem (e.g. via the websites of universities and research institutes) should be in the focus. For the analysis of textual content, several approaches may be suitable. If researchers want to investigate a topic that can be adequately described using a set of keywords (e.g. specific technologies, standards, patent numbers, policy measures) a simple keyword search can be sufficient. In such a keyword search, firms can be identified that use these keywords on their websites. Smarter search strategies with additional filtering words and the like may be used to refine the results.

Recent developments in the field of natural language processing (NLP) (e.g. Mikolov *et al.*, 2011, 2013; Mikolov, Yih and Zweig, 2013), especially the ones involving artificial neural network language models, resulted an array of potentially valuable approaches to extract innovation related information from web scraped texts. A possible approach to predict a firm's innovation activity as outlined in Figure 12. 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) could be used to enhance the model.

Text mining methods based on neural networks and semantic topic models were also successfully applied in geographical information science (GIScience) to uncover social phenomena from geocoded unstructured text data. Resch, Usländer, & Havas (2018) for example, present an approach to assess the footprint of and the damage caused by natural disasters by combining machine learning techniques for semantic information extraction. They also showed that their approach can be used to identify relevant semantic topics without a priori knowledge. Their methodology may be applicable to detect and monitor the diffusion of technology, for example.

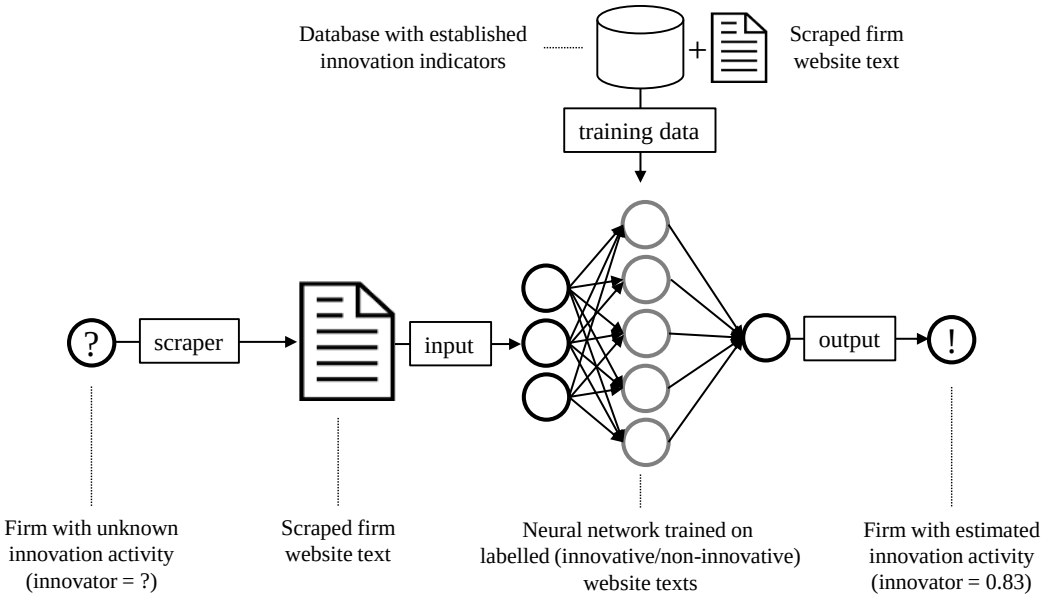


Figure 12. Proposed artificial neural network based innovation prediction model.

In this paper, we also showed that networks of hyperlink connected firms can be extracted from the web using ARGUS. Given that the appropriate metadata is available, specific regional and sectoral firm networks could be examined. Figure 13 maps such an exemplary network of firms that was scraped using ARGUS. Social network analysis offers an extensive set of widely adapted techniques for analyzing such networks in a quantitative manner (see e.g. Scott & Carrington, 2011). Future research could aim to find regularities in the structure of firm 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. Rammer, Kinne, & Blind, 2019).

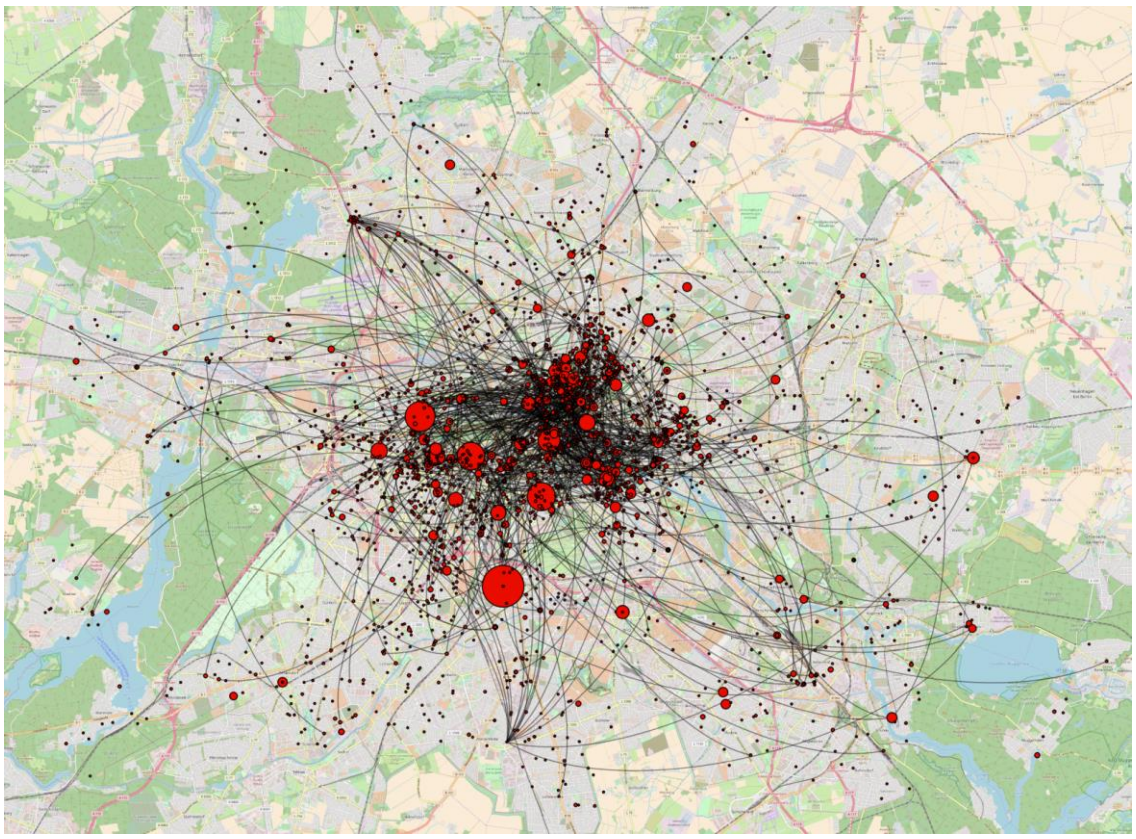


Figure 13. Exemplary map of hyperlink connections between software firms based in Berlin, Germany, generated with the approach developed in this paper. Basemap: OpenStreetMap

References

Ackland, R., Gibson, R., Lusoli, W., & Ward, S. (2010). Engaging With the Public? Assessing the Online Presence and Communication Practices of the Nanotechnology Industry. *Social Science Computer Review*, *28*(4), 443–465.

Acs, Z. J., Anselin, L., & Varga, A. (2002a). Patents and innovation counts as measures of regional production of new knowledge. *Research Policy*, *31*(7), 1069–1085. https://doi.org/10.1016/S0048-7333(01)00184-6

Acs, Z. J., Anselin, L., & Varga, A. (2002b). Patents and Innovation Counts as Measures of Regional Production of New Knowledge. *Research Policy*, *31*(7), 1069–1085. https://doi.org/10.1016/S0048-7333(01)00184-6

Archibugi, D., & Pianta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–468. https://doi.org/10.1016/0166-4972(96)00031-4

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: a pilot web-based study on graphene firms. *Scientometrics*, *95*(3), 1189–1207.

Arzaghi, M., & Henderson, J. V. (2008). Networking off Madison Avenue. *Review of Economic Studies*, *75*(4), 1011–1038. https://doi.org/10.1111/j.1467-937X.2008.00499.x

Askitas, N., & Zimmermann, K. F. (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

Basole, R. C., Huhtamäki, J., Still, K., & Russell, M. G. (2016). Visual decision support for business ecosystem analysis. *Expert Systems with Applications*, *65*(August), 271–282. https://doi.org/10.1016/j.eswa.2016.08.041

Basole, R. C., Russell, M. G., Huhtamäki, J., Rubens, N., Still, K., & Park, H. (2015). Understanding business ecosystem dynamics: A data-driven approach. *ACM Transactions on Management Information Systems*, *6*(2). https://doi.org/10.1145/2724730

Beaudry, C., Héroux-Vaillancourt, M., & Rietsch, C. (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* (pp. 1–25).

Behrens, V., Hünermund, P., Leitner, S. M., Licht, G., & Peters, B. (2018). *Investigating the Impact of the Innovation Union: State of Implementation and Direct Impact Assessment*. Maastricht.

Bersch, J., Gottschalk, S., Müller, B., & Niefert, M. (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, & TÜV Rheinland. (2016). *Broadband Atlas*. Berlin. Retrieved from https://www.bmvi.de/DE/Themen/Digitales/Breitbandausbau/Breitbandatlas-Karte/start.html

Carlino, G., & Kerr, W. R. (2015). Agglomeration and Innovation. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.), *Handbook of Regional and Urban Economics* (Vol. 5, pp. 349–404). Amsterdam: Elsevier North-Holland. https://doi.org/10.1016/B978-0-444-59517-1.00006-4

Catalini, C. (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, R. (1996). Core competencies and the strategic management of R&D. *R&D Management*, *26*(4), 345–355. https://doi.org/10.1111/j.1467-9310.1996.tb00970.x

Danilak, M. (2015). langdetect. Retrieved from https://pypi.org/project/langdetect/

Eurostat. (2018). EUROSTAT. Retrieved July 18, 2018, from 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

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

Getis, A. (2009). Spatial Weights Matrices. *Geographical Analysis*, *41*(4), 404–410. Retrieved from https://www.seas.upenn.edu/~ese502/lab-content/extra_materials/SPATIAL WEIGHT MATRICES.pdf

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Grentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as Data* (NBER Working Paper Series No. 23276). Cambridge, Massachusetts.

Griliches, Z. (1990). *Patent statistics as economic indicators: A survey* (NBER working paper No. 3301). *NBER working paper*. Cambridge, Massachusetts.

Jang, S., Kim, J., & von Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research*, *78*(June), 143–154. https://doi.org/10.1016/j.jbusres.2017.05.017

Katz, J. S., & Cothey, V. (2006). Web Indicators for Compley Innovation Systems. *Research Evaluation*, *45*(5), 893–909. https://doi.org/10.1016/j.respol.2006.03.007

Kerr, W. R., Duranton, G., Glaeser, E., & Henderson, V. (2014). Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics*, *96*(3).

Kim, J., Hwang, M., Jeong, D.-H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*, *39*(16), 12618–12625. https://doi.org/https://doi.org/10.1016/j.eswa.2012.05.021

Kinne, J. (2018). ARGUS - An Automated Robot for Generic Universal Scraping. Mannheim: Centre for European Economic Research. https://doi.org/10.1109/LPT.2009.2020494

Kleinknecht, A., & Reijnen, J. O. N. (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

Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). The Non-Trivial Choice between Innovation Indicators. *Economics of Innovation and New Technology*, *11*(2), 109–121. https://doi.org/10.1080/10438590210899

Krzywinski, M., & Altman, N. (2013). Points of significance: Significance, P values and t-tests. *Nature Methods*, *10*(11), 1041–1042. https://doi.org/10.1038/nmeth.2698

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, NY: McGraw-Hill Irwin.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.1162/153244303322533223

Mikolov, T., Deoras, A., Povey, D., Burget, L., & Cernocky, J. (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, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of NAACL-HLT* (pp. 746–751). https://doi.org/10.3109/10826089109058901

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Cambridge, Massachusetts: Academic Press.

Nagaoka, S., Motohashi, K., & Goto, A. (2010). Patent Statistics as an Innovation Indicator. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of Economics of Innovation* (Vol. 2, pp. 1083–1127).

Nathan, M., & Rosso, A. (2017). *Innovative Events* (Centro Studi Luca d'Agliano Development Studies Working Paper No. 429). Retrieved from https://ssrn.com/abstract=3085935

Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy*, *38*(6), 994–1005. https://doi.org/10.1016/j.respol.2009.01.023

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

OECD. (2017). *Broadband Portal*. Paris. Retrieved from www.oecd.org/sti/broadband/oecdbroadbandportal.htm

OECD, & Eurostat. (2018). *Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation* (4th ed.). Luxembourg, Paris: OECD/eurostat. https://doi.org/10.1787/9789264304604-en

Rammer, C., Aschhoff, B., Doherr, T., Peters, B., & Schmidt, T. (2017). *Innovationsverhalten der deutschen Wirtschaft. Indikatorenbericht zur Innovationserhebung 2016*. Mannheim. Retrieved from http://ftp.zew.de/pub/zew-docs/mip/16/mip_2016.pdf

Rammer, C., Kinne, J., & Blind, K. (2019). Knowledge proximity and firm innovation: A microgeographic analysis for Berlin. *Urban Studies*. https://doi.org/10.1177/0042098018820241

Raymond, K., & Blockeel, H. (2000). Web Data Mining research: A survey. *SIGKDD Explorations*, *2*(1), 1–10. https://doi.org/10.1109/ICCIC.2010.5705856

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376. https://doi.org/10.1080/15230406.2017.1356242

Rubens, N., Still, K., Huhtamaki, J., & Russell, M. G. (2011). A network analysis of investment firms as resource routers in Chinese innovation ecosystem. *Journal of Software*, *6*(9), 1737–1745. https://doi.org/10.4304/jsw.6.9.1737-1745

Scott, J., & Carrington, P. J. (2011). *The SAGE Handbook of Social Network Analysis*. SAGE.

Scrapy Community. (2008). Scrapy. Scrapinghub Ltd. Retrieved from https://github.com/scrapy/scrapy

Shepherd, W. G., & Shepherd, J. M. (2003). *The Economics of Industrial Organization*. Long Grove, IL: Waveland Press Inc.

Squicciarini, M., & Criscuolo, C. (2013). *Measuring Patent Quality* (OECD Science, Technologya dn Industry Working Papers No. 2013/03). Paris. https://doi.org/http://dx.doi.org/10.1787/5k4522wkw1r8-en

Steiger, E., Resch, B., & Zipf, A. (2016). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographic Information Science*, *30*(9), 1694–1716.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). London: Pearson.

Xu, G., Wu, Y., Minshall, T., & Zhou, Y. (2018). Exploring innovation ecosystems across science, technology, and business: A case of 3D printing in China. *Technological Forecasting and Social Change*, *136*(June 2017), 208–221. https://doi.org/10.1016/j.techfore.2017.06.030

Youtie, J., Hicks, D., Shapira, P., & Horsley, T. (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), 981–995. https://doi.org/10.1080/09537325.2012.724163

ZEW. (2018). ZEW-FDZ. Retrieved August 20, 2009, from https://kooperationen.zew.de/en/zew-fdz/home.html

Appendix

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 A1. Sectors' NACE code ranges.

Variable	Broadband	Population density	Employees	Age	Legal form	Search quality
Broadband	-					
Population density	-0.44	-				
Employees	-0.01	0.02	-			
Age	0.04	-0.08	0.04	-		
Legal form	-0.09	0.10	-0.10	-0.10	-	
Search quality	0.06	-0.10	0.15	0.37	-0.10	-

n=2,108,104; p≤0.001 for all correlations.

Table A2. Correlation (Spearman's rho) table.

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de