Rammer, Christian; Es-Sadki, Nordine

Working Paper

Using big data for generating firm-level innovation indicators: A literature review

ZEW Discussion Papers, No. 22-007

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

Suggested Citation: Rammer, Christian; Es-Sadki, Nordine (2022) : Using big data for generating firm-level innovation indicators: A literature review, ZEW Discussion Papers, No. 22-007, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/251523

// NO.22-007 | 03/2022

DISCUSSION PAPER

// CHRISTIAN RAMMER AND NORDINE ES-SADKI

Using Big Data for Generating Firm-Level Innovation Indicators – A Literature Review

Leibniz Association

ZEW

Using Big Data for Generating Firm-level Innovation Indicators - a Literature Review

Christian Rammer[a] and Nordine Es-Sadki[b]

[a] *Department Economics of Innovation and Industrial Dynamics, ZEW - Leibniz Centre for European Economic Research, Germany*
rammer@zew.de

[b] *UNU-MERIT, Maastricht University, The Netherlands*
n.es-sadki@maastrichtuniversity.nl

February 2022

Abstract

Obtaining indicators on innovation activities of firms has been a challenge in economic research for a long time. The most frequently used indicators - R&D expenditure and patents - provide an incomplete picture as they represent inputs and throughputs in the innovation process. Output measurement of innovation has strongly been relying on survey data such as the Community Innovation Survey (CIS), but suffers from several shortcomings typical to sample surveys, including incomplete coverage of the firm sector, low timeliness and limited comparability across industries and firms. The availability of big data sources has initiated new efforts to collect innovation data at the firm level. This paper discusses recent attempts of using digital big data sources on firms for generating firm-level innovation indicators, including Websites and social media. It summarises main challenges when using big data and proposes avenues for future research.

JEL-Classification: O30, C81

Keywords: Big data, innovation indicators, CIS, literature review

Corresponding author: Christian Rammer
Department Economics of Innovation and Industrial Dynamics
ZEW - Leibniz Centre for European Economic Research
L 7, 1
68161 Mannheim, Germany
Phone +49 621 1235 184;
Email: rammer@zew.de

Acknowledgements: The authors gratefully acknowledge support by the Statistical Office of the European Commission (Eurostat, Service Contract No. 2020.0163).

1. Introduction

Measuring innovation in firms for analysing the determinants and impacts of innovation activities has long been resting on R&D and patent data (see Cohen 2010, Griliches 1984) since these data are readily available from firms' financial accounts (R&D) or from publicly accessible data bases (patents). Both measures give a somewhat biased picture of firms' innovation activities, however, as not all innovations are based on own R&D efforts (Rammer et al. 2009, Som 2012) or patenting, and not all patents result in innovations (Griliches 2007). For fully capturing innovation in firms it would be required to measure innovation output, i.e., the introduction of new or improved products in the market or of new or improved processes in the firm (OECD and Eurostat 2018),

Using existing data for output measurement of innovation at the firm level turned out to be challenging. Early approaches using literature-based output indicators obtained from trade journals (Kleinknecht et al. 1993) or expert-based lists of major innovations (Geroski et al. 1997) suffered from a bias towards product innovation and economically important innovations. Dedicated firm surveys on innovation have been developed as an alternative approach for measuring innovation. Building upon first experiences from Germany (Meyer-Krahmer 1984) and Italy (Archibugi et al. 1987), the European Commission launched a large-scale innovation survey in 1992, the so-called Community Innovation Survey (CIS). This survey has become part of business enterprise statistics and is currently conducted in almost all European countries (Arundel and Smith 2013). Many non-European countries also regularly run surveys that collect information on innovation activities of firms, often following the CIS model, including Latin American countries (Castellacci and Natera 2012), Asian countries (Hong et al. 2012) as well as Canada, the United States, Australia and New Zealand.

Firm-level surveys on innovation come with some shortcomings, however. First, innovation survey data has been criticised on low international comparability due to differences in questionnaire design, sampling, and survey methods (Archibugi and Pianta 1996, Kleinknecht et al. 2002, Mortenson 2008). Secondly, innovation surveys are usually confined to a subset of the business enterprise sectors, often excluding very small firms or certain industries, hence failing to produce innovation data for the entire economy (Archibugi and Pianta 1996, Tether 2002, Cirera and Muzi 2020). Thirdly, innovation surveys apply definitions of innovation

which assume that all survey respondents understand in the same way, even though the interpretation of what is new or significantly improved to the firm is subjective (Arundel and Smith 2013, Cirera and Muzi 2020). In addition, the use of a uniform definition of innovation makes it difficult to represent special features of innovation in certain industries, limiting the comparability of innovation data across industries. Fourthly, innovation surveys are usually designed as sample surveys and are subject to unit and item non-response which may limit the statistical reliability of the results. Finally, conducting innovation surveys usually takes a significant amount of time (Kinne and Lenz 2021), resulting in a substantial time lag between the reference year of the data and the time the data is published. In addition, innovation surveys usually apply a subject-based approach (i.e., measuring innovation at the level of the firm and not for individual innovations) and provide little information on specific innovations or innovation related to certain newly emerging technologies or market trends.

Big data sources have the potential to overcome some of these shortcomings of innovation surveys and may offer a more complete picture of innovation in firms (Kinne and Axenbeck 2018, 2020). Key sources of big data that can be used for measuring innovation at the firm level include firm websites and social media activities of firms, but also other digital sources such as proprietary data sources, media reports, job offerings or online platforms. While none of these sources is devoted to report on innovation in firms, they may contain information that is related to activities of and events in firms that are linked to innovation (see Arora et al. 2016). Developing big data analytics promises to identify and extract this information.

The aim to this paper is to review empirical studies that used big data sources for measuring innovation in firms and to derive conclusions on how to exploit these sources for improving the production of firm-level innovation data, particularly compared to traditional approaches such as innovation surveys. The paper aims at complementing prior surveys on the use of big data in innovation research which focused on text mining techniques to analyse innovation management practices in firms (see Anton et al. 2020). The term big data is used in this paper to denote an information source that is digitally available and extensively covers a certain subject area, while it does not provide structured data that could be easily used to derive indicators but requires data mining techniques to extract useful information (in case of our paper: on innovation). The paper does not provide a complete literature survey but focuses on selected studies that represent typical approaches for using big data to generate innovation indicators.

Based on the experience made by the existing studies, we identify advantages and limitations of the use of big data and derive recommendations on good practice in the exploitation of digitalised big data for producing innovation indicators. The paper finally develops a proposal for future research on using big data for extending the reporting on innovation in firms, by producing more timely data, covering a larger set of industries and size classes, allowing for a more detailed geographical breakdown and providing details on innovation that go beyond the indicators collected in traditional innovation surveys such as the CIS.

The next section of the paper summarises existing studies. Section 3 discusses advantages and limitations of big data sources, and section 4 concludes with a research agenda for big data use and innovation indicators.

2. Existing Studies

Over the past ten years, an increasing number of studies explored the potential of publicly available big data to produce indicators on innovation activities of firms, following the general uptake in the use of big data analytics in research (Suominen and Hajukhani 2021). Some of these studies focussed on innovation in specific fields of technology while others looked at innovation by different types of actors, including firms, universities, public research organisations, private households and individuals. Antons et al. (2020) analysed 124 research articles that used text mining techniques for analysing innovation-related topics, mostly relying on other research papers and patent data. They identified several research areas in innovation management that could benefit from text mining techniques. Some are related to CIS type innovation indicators including data on innovation results based on product announcements or the detection of novelties based on bibliometric analysis of innovation hypes.

This section focusses on works that used big data from digitalised, publicly accessible sources for measuring innovation in firms. From the large number of such studies, we selected those that were among the first to develop and apply a certain methodological approach. For each publication, we summarise the data base employed, the empirical approach used, and the results that were obtained. Table 1 summarises main characteristics of each study. Five types of studies are distinguished: case studies, large-scale scraping of company websites (web crawling), analysis of social media, analysis of company reports and financial accounts, and the use of other digital sources on firms.

2.1. Case studies

Among the first case studies that examined the potential of digitalised big data sources on firms for analysing innovation activities of firms were Youtie et al. (2012) who examined current and archived website data of small and medium-sized firms (SMEs) from the United States active in the field of nanotechnology in order to identify the transition from discovery to commercialization. Based on a manual screening approach that started with 358 firms, the study finally used website data from a sample of 30 SMEs which was analysed using key word search an expert examination. The key innovation indicators examined included the type of product development activities in nanotechnology and the financing sources obtained for carrying out these activities. The entire process turned out to be very time-consuming, requiring about three months. The authors concluded that website information was useful for explore nanotechnology SMEs transitions from discovery to commercialisation and for understanding how transitions vary by SME characteristics, technology and market sectors. The analysis of smaller firms was more manageable since these firms tended to have smaller websites.

Other early case studies using a similar approach as Youtie et al. (2012) include Libaers et al. (2010) who examined keyword occurrence in company websites from a cross-industry sample of small and medium-size firms to identify commercialization-focused business models among highly-innovative firms, and Kim (2012) who analysed nanotechnology websites of different organisations for analysing the relationship between universities, government research labs and firms.

Among the first studies on big data sources and innovation in firms that used web crawling techniques for extracting website information was the study by Arora et al. (2013), building upon the findings of Youtie et al. (2012). They employ a web content analysis method to examine the activities of 20 SMEs in the US, UK and China related to the commercialisation of emerging graphene technologies. Based on a key word search, different application areas in the field of graphene technology were identified. Based on these search results, firms were classified into three groups according to the type of innovation activities: focus on product development, focus on materials development and focus on integration into existing product portfolios.

Gök et al. (2015) built upon the two case studies presented above, but extended both the sample size and the scope of innovation indicators. They used web mining to explore the R&D activities of 296 UK-based green goods SMEs. They find that website data offers additional insights when compared with other traditional research methods, such as patent and publication analysis. They examined the strengths and limitations of firm innovation web mining in terms of a wide range of data quality dimensions, including accuracy, completeness, currency, quantity, flexibility and accessibility. They find that, in contrast to only examining conventional data sources, companies in their sample, report more often that they undertake R&D activities on their website. They conclude that websites are a useful complement and may offer new insights not easily obtained from other sources. They also stress that specific technical skills are required, and transparency is needed about the methodological choices made. The same data source was used to identify collaboration of SMEs with universities and governments (see Li et al. 2018). A summary of the methodologies used by Gök et al. (2015), Li et al. (2018), Arora et al. (2013) and Youtie et al. (2012) can be found in Arora et al. (2016).

Another interesting case study is Rietsch et al. (2016) since they combine a website-based big data analysis with a conventional survey, allowing to analyse the consistency of results found by the two methods. Their exploratory study is based on data from 89 Canadian nanotechnology firms and looks at four groups of innovation indicators: R&D activities, use of intellectual property rights, collaboration with other organisations, and obtaining external financing. The validation of the web mining results with those from a classic questionnaire-based survey shows a significant positive correlation for all four indicators. The highest correlation coefficient is found for IP, and the lowest for collaboration and external financing. The authors conclude that some of the data extracted by the web mining technique can be used as proxy for specific variables obtained from more classical methods.

2.2. Large-scale web scraping

Among the studies that analyse websites of a large number of firms using web crawling and automated text analysis techniques in order to derive innovation indicators are two that are of particular relevance for the CIS; Kinne and Lenz (2019, 2021) and Daas and van der Doef (2020, 2021). Both studies build upon the same basic methodology. They use a sample of firms that participated in the CIS and for which data on the innovation status is available. This data is used as training data. For all firms in the CIS sample, information from the firms' websites is extracted and transferred into a text data base. Then a model is developed that uses this

text data base in order to predict the known innovation status of the firm. The model is designed in a way that its results show a very high fit with the innovation status of the firms. This model is then applied on the text of websites of firms which did not participate in the CIS in order to predict the innovation status of these non-CIS firms.[1] Since the CIS covers only a small fraction of the entire firm population, this methodology allows to derive innovation indicators for sectors and size classes not covered by the CIS. It also allows to derive innovation indicators for firms in the CIS core population that did not participate in the CIS.

The main difference between the two studies is the way the text on websites has been prepared, the type of model used, and the innovation indicator considered. Kinne and Lenz (2019, 2021) use a web scraping approach described in Kinne and Axenbeck (2018, 2020) and Kinne and Resch (2018) and developed a deep neural network for analysing website content. The text analysis rests on a dictionary of all words that occurred on websites as long as the document frequency is between 1.5% and 65% (i.e. very rare words and very common words are not considered), producing 6,144 different words. The innovation indicator analysed was whether a firm has introduced a new or improved product (product innovator). For training data, they used the German CIS, but considered only firms that were product innovators in three consecutive survey years (2015 to 2017) and firms that were no product innovators in all three survey years (exploiting the fact that the German CIS is a panel survey conducted every year). This choice was made in order to provide a clear cut between innovators and non-innovators. The focus on product innovation was motivated by the fact that firms are more likely to disclose information about product innovation than about process innovation. The accuracy of the model was tested with a sub-sample of the training data that was not used for training but put aside for testing. The result of the model is an 'innovation probability indicator' ranging between 0 and 1 and indicating how likely a firm is a product innovator based on the information contained on the firm's website. In order to compare the results, this probability has been normalised in a way that the average probability for the firms in the training data is the same as the average share of product innovators among the firms in the training data.

[1] A similar methodology was used by Mirończuk and Protasiewicz (2016). This study did not rely on innovation data reported by firms, but on a manual labelling of websites of 2,747 firms in terms of whether the website text indicated innovative or not innovative firms. The study does not provide any details on how the labelling was done.

Daas and van der Doef (2020, 2021), based on data from the Dutch CIS, perform a text analysis of websites that produced a database of 584 stemmed words.[2] This data base was analysed using logistic regression technique, since this method turned out to produce the best model fit (out of 11 methods tested). Logistic regression implies that the result of the model, i.e. whether a website content is classified as indicating an innovative or not innovative firms, primarily rests on the correlation of word occurrences in the website text and the innovation status of the firm. Daas and van der Doef (2021: 13) provide a list of the word stems with highest positive and negative coefficients. The innovation indicator used is whether a firm has introduced a new or improved product or process (innovator). The accuracy of the model was not only tested with a sub-sample of the training data, but also by manually checking the results for out-of-sample firms. This external validation was done for 933 start-ups (with a 98% confirmation rate) and for 1,000 other firms with less than 10 employees (with a 95% rate of correctly classified firms).[3] Daas and van der Doef (2021) also analysed the accuracy of the model when using different time lags between the CIS reference year and the time of website scraping and found a significant decline in accuracy for greater lags.

Both studies used the results of the models to predict the innovation status of firms not surveyed in the CIS, finding interesting results in terms of industries, size classes and regions. With respect to industries, Kinne and Lenz (2019, 2021) show that the website analysis produces a lower share of product innovators compared to CIS results, except for ICT services (NACE 61-63), wholesale trade (NACE 46) and consulting (NACE 70.2). Among the industries not covered by the German CIS, only one (management services, NACE 70.1) show an above-average product innovator share. All other industries report rather low shares, particularly construction (NACE 41-43) and health & social services (NACE 86-88).

[2] A word stem excludes grammatical features of a word, like a plural marker ('s' in English) or time markers for verbs (e.g. 'ed' in English for marking the imperfect at the end of a verb).
[3] No details are provided on how the innovation status of firms was established by the manual checking procedure.

Table 1 Selected characteristics of studies using big data sources for producing innovation indicators at the firm level

Study	Title	Big data source	No. of firms	Reference years	Innovation indicators	Methods	Validation of results	Main findings
Youtie, Hicks, Shapire, Horsley (2012)	Pathways from discovery to commercialisation: using web sources to track small and medium-sized firm strategies in emerging nanotechnologies	Websites (current and archived websites)	30 SMEs	1996 to 2010	• Product development	• Web crawling: not provided, probably manual copying of pre-selected webpages • Innovation indicators: key word search and expert examination	• Expert assessment	• Website information proved to be useful to explore nanotechnology SMEs transitions from discovery to commercialisation and understand how transitions vary by SME characteristics, technology and market sectors
Arora, Youtie, Shapire, Gao, Ma (2013)	Entry strategies in an emerging technology: a pilot web-based study of graphene firms	Websites (current and archived websites)	20 SMEs	1996 to 2010	• Application areas in the field of novel graphene technologies	• Web crawling: Texas A&M HHAT project • Innovation indicators: key word search	• Expert assessment	• Website information produced three groups of graphene firms: focus on product development, focus on materials development, and focus on integration into existing product portfolios
Gök, Waterworth, Shapira (2015)	Use of web mining in studying innovation	Websites (current and archived websites)	296 SMEs	2012 for current websites, 2004-2011 for archived websites	• R&D activities	• Web crawling: IBM Content Analytics (ICA) • Innovation indicators: key word search	• Expert assessment	• Higher share of R&D active firms based on website data compared to data on R&D expenditure or receipt of R&D grants • Website-based R&D indicator not correlated with patents and publications by the same firms • Website data can complement traditional sources of innovation data, but not substitute them
Rietsch, Beaudry, Héroux-Vaillancourt (2016)	Validation of a web mining technique to measure innovation in the Cana-	Websites	89 firms, of which 79 provided sufficient website information	not provided (probably 2015)	• R&D activity • Intellectual property activity • Collaboration	• Web crawling: Nutch • Innovation indicators: key word search	• Questionnaire-based survey	• Positive correlation between website-based indicators and indicators obtained from a questionnaire-based survey, with best results found for IP, and worst for collaboration and external financing

	dian nanotechnology-related community							• Web mining R&D indicator did correlate the most when firms were more likely to provide R&D services to third parties, and when they had a high share of R&D employees, but no correlation with number of R&D projects
Kinne, Lenz (2019, 2021)	Predicting Innovative Firms using Web Mining and Deep Learning	Websites (current websites)	Training data: 3,126 firms with usable website information (out of a sample of 4,481 enterprises from the German CIS 2016 sample that also participated in the two preceding surveys); total no. of analysed websites: 1.15 million	2012 to 2016 for innovation status, 2018 for website information	• Product innovation	• Web crawling: ARGUS (tool developed by the authors) • Innovation indicators: deep neural network, training data from CIS • Application of trained model to about 700,000 websites of firms in Germany	• Community Innovation Survey	• Comparing the product innovator predictions based on website information with patent statistics, CIS indicators, and regional innovation indicators shows that the predictions are plausible and produce consistent results. • The innovation indicator derived from website analysis can be used to produce innovation data that is more disaggregated in terms of industries, size classes, regions or other characteristics of firms (e.g. age).
Daas, van der Dreef (2020, 2021)	Using Website texts to detect Innovative Companies	Websites (current websites)	Training data: 4,765 firms with usable website information (out of a sample of 6,342 firms from the Dutch	2014 to 2016 for innovation status, not exactly given for website information	• Product or process innovation	• Web crawling: Python 3.7 (as well as Node.js, Selenium) • Innovation indicators: logistic regression of 584 stemmed words obtained from webpages on the probability to innovate, training data from CIS	• Community Innovation Survey • Expert assessment	• Manual validation of results for start-ups and for a random sample of 1,000 firms not contained in the Dutch CIS 2016 reveal a high accuracy of the model predictions. • Combining website information from several points in time provided the best model fit.

			CIS 2016 sample); total no. of analysed websites: 0.5 million	('freshly scraped'), probably 2017 to 2018		• Application of trained model to 466,523 websites of firms in the Netherlands with less than 10 employees		• Enlarging training data by manually classified websites improves model fit.
Nelhans (2020)	Data4Impact	Websites (current websites)	1,370 firms with usable website information (out of a sample of 2,331 firms with EU-funded projects in the field of health research that are documented in CORDIS)	2014 to 2016 for innovation status, not exactly given for website information ('freshly scraped'), probably 2017 to 2018	• Goods innovation • Other innovation • Use of intellectual property rights • Other innovation activity	• Web crawling: in-house developed tool • Innovation indicators: machine-learning models, training data from manual coding of websites	• None	• 30% of firms were classified as goods innovators, 15% showed evidence for service or process innovations. 6% of firms were classified as using IPRs, 14% as active in mergers and acquisitions. • 5% of firms were classified as having received private funding and 8% public funding. These low shares demonstrate that websites do not contain direct information on funding sources used by firms, since most firms in the sample will have received private funding, and all have received public funding from the EU. • The results are determined by the manual classification of websites, the accuracy of which is difficult to assess.
Breithaupt et al (2020)	Intangible Capital Indicators Based on Web Scraping of Social Media	Facebook, Kununu	1,539 for Facebook, 2,114 for Kununu (out of a gross sample of 8,278 firms)	2017 for Facebook, 2010 to 2018 for Kununu	• Number of Facebook Likes (indicator of brand equity) • Firm ranking on "company image" (indicator of brand equity) • Firm ranking on "on-the-job	• Google search on Facebook pages of firms and Kununu mentioning of firms • Innovation indicators: Download of relevant Facebook and Kununu pages and reading out the relevant information	• Community Innovation Survey	• Positive and statistically significant relationship between survey-based data on marketing and training expenditure, and the respective information stemming from the social media platforms

					training/career development" (indicator of firm-specific human capital)			
Garechana, Río-Belver, Bildosola, Rodríguez Salvador (2017)	Effects of innovation management system standardization on firms: evidence from text mining annual reports	Company reports	12 firms	2002 to ? (probably 2014)	• R&D activity • Intellectual property activity • Collaboration	• Text mining: Vantage Point • Innovation indicators: emergence of words related to innovation, e.g. research, development, innovation, new technologies, new materials	• Comparison of pre-certification and post-certification patterns • Expert assessment	• Relevance of emergent technologies in the firms' innovative efforts increases and older technologies become less relevant after the adoption of innovation management standards • Terms that indicate a more open, more collaborative concept of innovation tend to emerge more frequently • Precision of text mining of innovation-related terms in company reports it assessed to be satisfactory
Gandin, Cozza (2019)	Can we predict firms' innovativeness? The identification of innovation performers in an Italian region through a supervised learning approach	Balance sheets from administrative records	2,688 firms (873 from CIS)	2011 and 2013	• Innovation activity (positive innovation expenditure)	• Random forest machine learning algorithm	• Community Innovation Survey • Patent data	• Main predictors of innovativeness are the firm's industry affiliation, turnover, and the share of intangibles in total assets • Model predicted 72% of firms with patent applications as being innovative

Source: authors' compilation

In terms of size classes, Daas and van der Doef (2020, 2021) show that firms with less than 2 working persons show a higher share of innovators (47%) than firms with 2 to less than 10 employees (35 to 40%). Kinne and Lenz (2019, 2021) show that the prediction results of website analysis for the share of product innovators are significantly lower than the reported shares from the CIS for firms with 500 or more employees, and slightly higher for firms with 20 to 99 employees. For firms with less than 5 employees and with 5 to 9 employees, the same product innovator share are predicted as for firms with 10 to 19 employees.

Daas and van der Doef (2021) demonstrate that website-based innovation indicators can be calculated at a highly disaggregated regional level (e.g. 4-level zip codes). Kinne and Lenz (2019, 2021) show that product innovator shares tend to be significantly higher in urban agglomerations compared to rural areas. Within urban agglomerations and larger cities, there are substantial differences at the local level. Data for Berlin reveal certain 'innovation hot spots', which tend to be more pronounced based on website information compared to survey data.

The methodology of using innovation survey data to train models for analysing big data sources has the main advantage that no manual classification of innovation variables is required. At the same time, only innovation indicators collected in innovation surveys can be used, i.e. this methodology does not allow to establish innovation indicators that complement the indicators from innovation surveys. The methodology's main advantage is to produce innovation data for firms not part of the sample of the innovation survey, including firms outside the survey's target population.

The website data produced by Kinne and Lenz (2019, 2021) was used to investigate other innovation-related data using website information. Krüger et al. (2020) analysed the interaction of innovative firms with other firms and other organisations based on hyperlinks found on the firms' websites. CIS data on innovation cooperation was used to validate the results, Mirtsch et al. (2021) used the website data to analyse the adoption of information security management system standards by firms.

In an ongoing research project, the approach of Kinne and Lenz is developed further to include information on changes of website content for predicting innovating firms.[4] The motivation for this research is that innovation is an inherently dynamic phenomenon. In order to

[4] See https://www.zew.de/PJ3421-1

qualify for an innovation, a new or improved product or process has to differ significantly from the firm's previous products and processes. One could hence expect that the changes that constitute an innovation should also be reflected in changes in the website content of a firm.

Using website scraping and automated text analysis for measuring innovation in firms has been used in many other studies. In the absence of verified data on the firms' innovations and innovation activities, many of these studies rely on simple text analyses, often based on manual coding of a sample of websites or by defining key words that indicate innovation. One example for this approach is the EU-funded project Data4Impact.[5] One part of this larger project was to analyse websites of firms that have received R&D funding from the EU in the field of health research in order to derive indicators on innovation output. The study (Nelhans 2020) developed a machine-learning model based on training data which were generated by manual labelling of website texts. The results are rather inconclusive. While all firms in the sample have received EU funding, the website analysis found only 8% of firms that received any type of public funding (EU, national or regional). The share of firms classified as goods innovator is rather low (30%), as is the share of firms with other innovations (15%), given that all firms are R&D performers and received public funding from the EU Horizon 2020 programme.

A similar exercise to Nelhans (2019) was done by Pukelis and Stanciauskas (2019) who analyse websites of 1,301 firms, using manually labelled training data. In the manual labelling process, the websites of 500 firms have been assigned into one group with innovation-related text on the website, and another one with no such text. No detail is provided how the manual classification was performed. As the results are of the analysis of websites is determined by the manual classification of websites, the accuracy of the result is difficult to assess. In another recent research, Ashouri et al. (2021) use website data of a sample of about 90,000 firms from selected manufacturing industries in the EU to identify firms with product innovation, product digitalisation, collaborations and the use of standards.

A general conclusion of these labelling-based big data techniques is that one needs high-quality training data, e.g. data on innovation reported by firms in traditional surveys or obtained from other 'objective' and representative sources, in order to perform website analyses that produce meaningful and comprehensible innovation indicators.

[5] See https://cordis.europa.eu/project/id/770531

2.3. Social media

Social media is another digital big data source that has been extensively used for various purposes (see Sloan and Quan-Haase 2017). With respect to innovation in firms, most research focuses on the use of social media by firms as part of their marketing and innovation activities, e.g. to obtain information on new user preferences (crowd sourcing) or to test innovative ideas with potential users (see Bhimani et al. 2021, Bruhn et al. 2012, Misirlis and Vlachopoulou 2018), or as a source of own big data analysis (see Niebel et al. 2019). Social media data are primarily used to assess their use by firms (see Arora et al., 2014), the impact on their performance (Coursaris et al. 2016, Chung et al. 2015, Tirunillai and Tellis 2012, Luo et al. 2013) or the link to human capital indicators (e.g. Gortmaker et al. 2021, Aguado et al. 2019, Chiang and Suen 2015, Zide et al. 2014, Ji et al. 2018, Pisano et al. 2017, Banerji and Reimer 2019), but only rarely to derive indicators on the innovation activities of firms. Most studies using social media for innovation indicator construction relate to the public perception of innovations or new technologies. For example, Veltri (2013) carried out a semantic analysis on 24,000 tweets from Twitter to understand the public perception of nanotechnology. Nakatsuji et al. (2006, 2009) used internet blogs to analyse 'innovative topics'. Albert et al. (2015) aimed to measure technology maturity based on the analysis of internet blogs.

With respect to innovation indicators for firms, the study by Breithaupt et al. (2020) is an interesting approach for using the potential of social media, though it is mainly explorative in nature. The main goal of the study is to generate indicators on two types of intangible assets that are related to innovation and may complement innovation data from surveys such as the CIS. One indicator dimension relates to economic competencies related to marketing and branding. Information from Facebook platform profiles of firms is used to derive the number of Facebook Likes a firm has received. This number is used as an indicator for the brand equity of a firm (see Coursaris et al. 2016). A second indicator of brand equity is derived from the German employer branding and review platform Kununu (which is similar to the US-based platform LinkedIn). On this platform, current and past employees can rank a company with respect to "company image". A second indicator dimension relates to firm-specific human capital. For this dimension, the ranking for "on-the-job training/career development" from the platform Kununu is used.

In order to validate the results, the indicators obtained from social media are compared to indicators on brand equity and firm-specific human capital reported by firms in the CIS. For this

purpose, a sample of firms from the German CIS 2016 is used. The German CIS 2016 collected data on marketing expenditure (i.e. expenditure on marketing research, advertising and other activities to increase brand equity of a firm) and on training of employees. Both expenditures correspond to the respective variables used in the harmonised data collection of the CIS 2018. The analyses showed a positive and statistically significant relationship between the survey-based indicators (expenditures on marketing and training) and the respective information stemming from the online platforms.

2.4. Company reports and financial accounts

Another big data source for deriving innovation indicators are company reports. Usually, firms provide detailed information on their business activities in such report, with a special emphasis on the developments that took place within the reporting year and that are relevant for assessing the future prospects of the firm by investors and the wider public. An interesting case study on this big data source is Garechana et al. (2017). They use company reports in order to identify changes in firms' innovation-related activities, following the adoption of a certified innovation management standard (UNE 166002). In the case study, company reports of 12 firms were analysed using text mining techniques. Garechana et al. (2017) find that the relevance of emergent technologies in the firms' innovative efforts increases and older technologies become less relevant after the adoption of innovation management standards. In addition, terms that indicate a more open, more collaborative concept of innovation tend to emerge more frequently.

This research demonstrates that company reports could be used for deriving innovation indicators on firms, complementing the widespread use of company reports as a source on other areas of firm activities and performance (see Back et al. 2001). However, there are several limitation to this type of big data as a source of innovation indicators. First, company reports are usually highly standardised, following the reporting standards required by authorities. This clearly limits the usefulness for analysing topics that are not part of the standard reporting content, which also applies to the topic of innovation. Secondly, company reports are available only for a small fraction of firms, i.e. very large firms and publicly listed firms. In most countries, there are no company reports for the vast majority of SMEs. This data source is hence not suited for deriving innovation indicators for the entire firm population. Thirdly, there are clear incentives for firms to highlight activities and results that are valued positively

by investors and the wider public. Since innovation is often regarded as an activity that positively contributes to future firm performance, there is a risk of over-reporting innovation, e.g. by describing changes as innovation that do not fulfil the innovation criteria of the Oslo Manual. This bias has to be taken into account when analysing this data source.

In a similar approach, Gandin and Cozza (2019) attempted to predict the innovation status of firms based on an analysis of publicly available financial account data. They employed a machine learning approach and used training data from the Italian CIS and the Italian R&D survey (reference years 2012 and 2014) for a sample of 935 firms from the region of Friuli-Venezia Giulia (873 observations coming from the CIS). The innovation indicator used was an indicator for having positive innovation (or R&D) expenditure or no innovation/R&D expenditure. The trained model was then used to predict the innovation status of a sample of 2,688 firms for which financial account data was available. In addition to financial account information such as turnover, number of employees, tangible and intangible assets, long and short term debt, and profits, the authors also assigned firms to technology classes based on industry codes. By using financial account data, the authors were able to predict the innovation status for firms outside the CIS target population. In addition, the results were also used to predict the patent status of a firm, i.e. whether it applied for a patent in one of the two reference years. The model predicted 72% of patent applicants as innovative.

2.5. Other big data sources

There are further big data sources that can be used to derive innovation indicators for the firm sector, though all are subject to specific limitations.

Data on funding activities of public authorities includes information on innovation activities of firms that received public support through government programmes, such as grant programmes to co-finance R&D and other innovation projects, loan programmes for investment in new equipment, or programmes that proved consulting services for innovation or support human capital development activities related to innovation. Such data often include a description of the activity that receives funding that can be used to obtain more detailed information on innovation activities (e.g. type of technology, targeted markets, cooperation). A typical example for such a data source is the Cordis data base of the EU. The limitation of this source is obvious as it is available for publicly funded innovation activities only. Data from the CIS

show that only a small fraction of firms with innovation activities receive public support. The results obtained from that source are hence highly biased.

Intellectual property rights data such as patents or trademarks have been widely used as innovation indicators for a long time, though its limitations have been pointed out early (see Pavitt 1985, Griliches 1998). Patents represent inventions, and not all inventions are transferred into innovation, while most innovations do not rely on patents (Arundel and Kabla 1998, Kleinknecht et al. 2012). In addition, only a fraction of new knowledge relevant for innovation is patented. In services, only specific types of new knowledge can be patented. As a result, patents are a highly biased indicator of innovative activity. Trade mark data have also been used as a source for innovation indicators (Mendonça et al. 2004), particularly in the service sector (Schmoch 2003, Schmoch and Gauch 2009). While trademarks are more widely used by firms, their link to innovation is less obvious than for patent data as trademarks are basically a mechanism to protect investment in brand equity, which may be related to innovation, but can also be used for brands that do not relate to innovations, but to old products (see Crass 2014a,b).

Bibliometric data is primarily used to analyse the scientific output of researchers. It allows, among others, to analyse the relevance of different research topics, co-operation in research, and changes in research output over time. A small part of scientific publications is produced by researchers in firms. Among all firms with R&D activities, only a small fraction do publish (Krieger et al. 2021). For this reason, bibliometric data are of very limited use for producing innovation indicators that cover a larger part of the firm population.

Data from trade journals (including industry and technical journals) often report about products that were newly developed and introduced on the market by firms. Similarly, catalogues of trade fairs and other publications related to trade fairs also include such information. This source has early been used to generate so-called 'literature-based innovation indicators' (Coombs et al. 1996). The limitations of this source include a strong bias towards industries for which trade journals exist and which organised trade fairs. This is often the case for certain manufacturing industries, but less for services. In addition, only specific types of innovations will be covered in trade journals and trade affairs, with a strong focus on product innovations and innovations with a higher (expected) impact on markets. Process innovation are hardly captured by this source.

Similarly to trade journals and trade fair information, press releases of firms as well as reports about firms in newspapers and magazines, including online media, are another source for identifying innovations of firms (see Kahn 2018) or newly founded innovative firms (see Von Bloh et al. 2020). This source suffers from similar limitations as trade journals and trade fair catalogues as the focus will be on more important product innovation, while neglecting incremental product innovation and process innovation. However, coverage of service innovation may be better using press releases as compared to trade journals. There are only few, rather experimental works using press releases as innovation indicator, e.g. Ikeuchi (2017).

Another potential source to inform about innovation activities of firms are job announcements by firms, which are nowadays usually published online on the website of a firm or on employee platforms such as LinkedIn (Hamilton and Davidson 2018). Very few studies have yet has used this source for producing innovation indicators (see Apatsidis et al. 2021 for an application based on data from the Stack Overflow platform), although the source should be less subject to limitations than the others listed above. First, most firms from all industries, including SMEs and firms from non-innovative industries, are likely to announce their new job openings. This is particularly true for more advanced industrial countries with an increase in labour shortage due to demographic change. At the same time, job announcements usually include details on the tasks to be performed, which often can be linked to innovation activities (e.g. if a reference is made to R&D, new product development, marketing of new products and services, implementation of new technology and equipment, etc.).

There are also several proprietary datasets that have been used for generating innovation-related indicators. For example, CrunchBase is a proprietary database of start-ups has emerged in recent years as a potential collection of innovation-relevant data and a primary source of data for investors (Dalle et al. 2017). Crunchbase gathers data on businesses, including founding year, funding raised, funding rounds, number of investors, acquisitions, etc. The OECD study by Dalle et al. (2017) benchmarked CrunchBase's coverage against the OECD Entrepreneurship Financing Database and other sources (e.g. VentureXpert or PwC). They conclude that patterns across years and sectors were similar, suggesting that coverage is quite comprehensive, particularly for start-ups in the United States. Crunchbase has also been used in a study exploring innovation ecosystems in the UK and US (Kemeny et al. 2017) and to analyse the effects of environmental policies on innovations with environmental benefits by start-ups (Cojoianu et al. 2020). Another group of frequently used proprietary big data sources

are company databases that contain basic economic and ownership information on a substantial part of the business enterprise sector, such as the Orbis database of Bureau van Dijk. Applications with respect to innovation indicators have been rare, however, since these databases do not contain direct information on innovation (except from patent information). Some authors used industry classifications as a proxy for innovativeness (see Ruhrmann et al. 2021).

3. Advantages and Limitations of Big Data as a Source for Innovation Indicators

The existing studies that use (digitalised) big data for collecting information about innovation in firms reveal that this source has a number of advantages over traditional data collection methods based on firm surveys using standardised questionnaires, such as the CIS:

3.1. Advantages over traditional data collection methods

The existing studies demonstrate a number of advantages of big data over traditional data collection methods based on firm surveys using standardised questionnaires, such as the CIS, which relate to timeliness, frequency of data production, cost, completeness, accessibility, flexibility, and content (see also Gök et al. 2015).

Timeliness

The distance between the reference time to which data refer to and the time of data collection is extremely short and can even be zero if big data sources are being analysed in real-time. Survey-based data usually have a time lag of half a year or more. In case the results of the big data analysis are reported quickly, big data is an excellent source for timely data on innovation.

Frequency

Big data analysis can be repeated in short intervals at almost no cost, provided that data updates are easily available. In case the data source is updated continuously, indicators based on big data could even be produced in continuous real time. In addition, frequent updating allows to set-up panel data with very short time intervals, whereas survey data on innovation usually work with intervals of one or two years.

Cost

The main costs of big data analysis relate to the preparation of the raw data from the big data source and to the development of a code. These costs are typically much smaller than survey costs. Repeating big data analysis comes at almost no additional cost.

Completeness

Depending on the nature of the big data source, big data analysis can provide comprehensive data on the entire firm population, offering census-like data that are not subject to sampling errors or restricted to certain sections of the firm sector (e.g. certain size classes, certain industries).

Accessibility

In case big data analysis rest on publicly available data (e.g. websites of organisations, social media, open source publications, other public data), the results can be published at the level of individual firms. This is usually not possible for survey-based data due to confidentiality regulation. Big data sources allow to establish micro-level data bases. It is also possible to link micro-level data from different big data sources, increasing the analytical power of the micro data base.

Flexibility

Provided that big data sources contain detailed and reliable information on innovation-related topics, big data analysis can be used to investigate new themes and produce innovation indicators on emerging topics, e.g. for new fields of technology (nanotechnology, artificial intelligence, etc.) or for new types of organising innovation processes in firms (cooperation, open innovation, user innovation, etc.). In traditional innovation surveys, taking up a new topics often requires time-consuming development work and testing for new questions and indicators.

Content

Many big data sources such as websites, social media, press releases or media reports provide information that refers to more downstream activities in the innovation process, including new product launches and commercialisation activities related to new products (see Antons et al. 2020). Big data could hence complement well-established input and throughput indicators on innovation such as R&D and patents on the output side. This usage of big data would continue the tradition of literature-based innovation indicators (Coombs et al. 1996, Kleinknecht et al.

1993, Geroski et al. 1997) while offering a more complete coverage of innovation output since most digital big data sources cover a much larger fraction of the firm population in terms of industries and countries than traditional literature such as trade journals or catalogues of industry fairs do. Using text mining techniques to analyse these data sources allows to detect innovation-related indicators such as novel techniques in scientific articles and patents and the identification of innovation and media hypes (Antons et. al 2020). One of its biggest potential is to measure the diffusion of innovations, but there are to our knowledge very few studies that have used big data to analyse technology diffusion. One example is Yu et al. (2020) who have analysed the diffusion of digital printing technology using social media based data analytics along with data mining and traditional statistical modelling.

3.2. Limitations of big data for producing innovation indicators

The advantages of big data sources are partly counterbalanced by a number of limitations that restrict the value of big data for producing reliable and useful innovation indicators, including biased information, limited coverage, lack of accuracy and consistency, a biased coverage of the innovation process, varying currency, language bias, and the need for interpretation.

Biased information

Information provided in big data sources is usually self-reported and deliberately selected, either by firms themselves or by others (e.g. in case of trade journals, social media or digital media). The motivation for posting or not posting information on websites and other digitalised big data sources may vary and is likely to be biased towards information that is supports a positive reception of the firm and its activities by the audience of digital sources, including investors, media, researchers, governments, and the general public. For social media

Limited coverage

Big data sources many not cover all firms, and the firms not covered are usually not a random sample of all firms but show some systematic differences to firms which are represented in a big data source. This is particularly true for social media data which are available only for a small fraction of all firms (see Breithaupt et al. 2020), with a bias towards larger firms and firms supplying private households. In addition, the depth of information available may vary considerably across firms. While almost all firms run a website, some firms have a very sim-

ple internet presence with just one web page, while others operate hundreds of webpages, offering a great wealth of information (see Kinne and Axenbeck 2021). Analysing such diverse data bases is likely to produce biased results.

Lack of accuracy and consistency

Comparing innovation-related information from big data sources is likely to suffer from different levels of accuracy and varying definitions of key terms. First, firms will apply their own definition of innovation, potentially leading to inconsistent results on innovation activities across firms. Secondly, comparability with data from innovation surveys based on the definitions proposed in the Oslo Manual will be limited as firms may understand innovation in a different way, e.g. considering only new-to-market innovations or R&D-based innovations, or focus on product innovation. Thirdly, firms may over-represent certain activities, for example, claiming new product developments that are perhaps neither new nor innovative.

Biased coverage of the innovation process

Information about innovation in most big data sources is likely to relate to product innovation, while only limited information is offered on process innovation. The reason is that most (digitalised) big data sources are used by firms to communicate with others. The primary target of communication are typically the users of their products who are informed about the offerings of the firm (see Kinne and Lenz 2021). There is significantly less reason for firms to inform others about their process innovation.

Varying currency

Although most big data sources look very timely, the information provided may refer to a point in time long ago and may be outdated at the time of analysis (Gök et al. 2015). Currency of information strongly depends on how frequently information is updated in a big data source, and which portions of the information are updated. Using panel information from big data sources can be very helpful in assessing the currency of the information available at the point of time the big data analysis is carried out.

Language bias

A challenge that has been widely pointed out in the literature, particularly in regard to web scraping and mining, is the challenge raised by language differences across countries (Klinger et al. 2018). These appear to have been at least partially overcome by advances in language

translation capacities, including through openly available Application Programming Interfaces (APIs) such as Google Translate. However, challenges remain, with the most advanced language processing tools still developed for the English language.

Need for interpretation

Big data sources are often difficult to interpret and the results derived from the sources are sensitive to methodological choices such as the search strategy used to extract information, the way the raw data are processed, the models used to analyse the data, and the methods employed to validate the results (see Gök et al. 2015). In case where external training data are available, many big data analysis rest on manually coding training data or on key word search. In both cases, the results are largely determined by the decision of the individuals labelling a sample of the big data source in terms of innovation-related content, or by the choice of key words. Most existing studies following this approach are very brief on providing details on how exactly the coding took place and how sensitive the results are on choices made in the coding process or when selecting key words.

4. Conclusions - Towards a Research Agenda for Big Data Use for Innovation Indicators

Big data undoubtedly offers a relevant new source for obtaining information about innovation in firms. It could extend the reporting on innovation in firms by producing more timely innovation indicators for all industries and size classes at a highly disaggregated regional level. Moreover it offers the opportunity for additional innovation indicators, e.g. on specific technologies. In order to exploit this source, the following three conditions should be met.

First, big data should cover the entire population of firms for which innovation indicators should be produced (*completeness*). If this condition cannot be met, methods should be employed to control for the resulting bias when deriving innovation indicators, e.g. by applying different weights for differently covered groups of firms. In case big data do not cover the entire population, the population represented in big data should not be biased but represent the population with respect to main population characteristics (e.g. size, age, industry, location of firms).

Secondly, the information available for each firm in a big data source should be of similar detail and not subject to biases across firms (*unbiased data*). This is particularly the case for

websites as content can vary across firms for various reasons. In case this condition cannot be met, correction factors for firms with insufficient detail on information should be applied. In order to establish correction factors, comparison with data from unbiased sources such as innovation surveys that are based on stratified random sampling should be used.

Thirdly, innovation indicators derived from big data analysis should be *valid and reliable*. Depending on the big data source, there can be various sources for limited validity and reliability, e.g. a focus on specific types of innovations or over-reporting of innovation. The validity of innovation indicators from big data sources should be analysed against the results of official innovation surveys for indicators that are included in both sources. If possible, models for analysing big data should be trained on data from official innovation surveys (see Kinne and Lenz 2019, 2021, Daas and van der Doef 2020, 2021). If no such validation is possible, verification of results based on expert assessments should be applied.

Fourth, in line with Antons et al. (2020) recommendations, innovation indicators derived from big data analysis should be *transparent*. The methodology should explain the texts being used, the sources, amount of data analysed, software and techniques used to run the analysis, and details about the algorithms employed and the choices made to fine-tune the algorithm. Transparent studies enable other scholars to replicate in for instance other countries and allows to increase the use of big data techniques to derive innovation indicators.

In order to utilise this potential of big data for extending the reporting on innovation in firms, research should observe a number of principles. First, the methodological conditions outlined above should be followed, including a coverage of the entire business enterprise sector and apply weighting and correction methods to adjust for biases in the data source. In order to achieve a high validity of big data results, models for analysing big data should be trained with data from firm surveys that apply standard definitions of innovation such as those of the Oslo Manual (see Kinne and Lenz 2019, 2021, Daas and van der Doef 2020, 2021, Gandin and Cozza 2019). The value of innovation indicators from big data analysis can substantially be increased if they can be linked to additional firm-level information. This includes basic information on the firm such as size, age, industry and location. Using international classification standards for industries and locations would allow to link the indicators to other business statistics and enhance the analytical potential of the data.

A main purpose of big data use should be to extend the coverage of traditional innovation surveys. A main value is to offer more timely and more disaggregated data that cover all parts of

the business enterprise sector. When training big data models on survey data, the analysis can be used to produce innovation indicators earlier than innovation surveys can do, for a much more detailed breakdown in terms of industries, size classes, regions and other firm characteristics (e.g. age) for the entire business enterprise sector. By using training data from innovation surveys, the big data results can be normalised in a way that indicators are directly comparable to survey results (see Kinne and Lenz 2019, 2021).

Big data analysis are particularly valuable for producing additional innovation indicators that go beyond the data collected through surveys, e.g. on the diffusion of specific technologies. For such applications, there are usually no questionnaire-based survey data available to train big data models. Instead, expert coding and key word search will have to be used. Analyses should focus on topics for which an unambiguous and unbiased identification of indicators is possible, e.g. if key words can be established that will be used by all firms in the same way. This has proved to work in certain fields of technology, e.g. for nanotechnology (see Kim 2012, Rietsch et al. 2016, Youtie et al. 2012, Veltrie 2013). Other areas for which big data can serve as a source for additional innovation indicators may include artificial intelligence (Kinne and Axenbeck 2020) or the impact of crises on innovation (see Kinne et al. 2020 for the case of Covid-19).

The use of (digital) big data for generating firm-level innovation indicators is still at its beginning, and much more methodological research is required in order to establish reliable and meaningful indicators. This paper presented some principles for future research on big data use to generate innovation indicators. Exchanging experiences among researchers and conducting more experimental research would be highly useful in this respect.

5. References

Aguado, D., J.C. Andrés, A.L. García Izquierdo, J. Rodríguez (2019), LinkedIn “big four”: job performance validation in the ICT sector, *Journal of Work and Organizational Psychology* 35(2), 53–64.

Albert, T., M.G. Moehrle, S. Meyer (2015), Technology maturity assessment based on blog analysis, *Technological Forecasting and Social Change* 92, 196–209.

Antons, D., E. Grünwald, P. Cichy, T.O. Salge (2020), The application of text mining methods in innovation research: current state, evolution patterns, and development priorities, *R&D Management* 50(3), 329–351.

Apatsidis, I., K. Georgiou, N. Mittas, L. Angelis (2021), A study of remote and on-site ICT labor market demand using job offers from Stack Overflow, *47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*.

Archibugi, D., S. Cesaratto, G. Sirilli (1987), Innovative activity, R&D and patenting: the evidence of the survey on innovation diffusion in Italy, *Science Technology Industry Review* 2, 135–190.

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

Arora, A., A.S. Arora, S. Palvia (2014), Social media index valuation: impact of technological, social, economic, and ethical dimension, *Journal of Promotion Management* 20(3), 328–344.

Arora, S.K., Y. Li, J. Youtie, P. Shapira (2016), Using the wayback machine to mine websites in the social sciences: a methodological resource, *Journal of the Association for Information Science and Technology* 67, 1904–1915.

Arundel, A., I. Kabla (1998), What percentage of innovations are patented? Empirical estimates for European firms, *Research Policy* 27(2), 127–141.

Arundel, A., K.H. Smith (2013), History of the Community Innovation Survey, in F. Gault (ed.), *Handbook of Innovation Indicators and Measurement*, Edward Elgar Publishing, 60–87.

Arundel, A., K. O'Brien, A. Torugsa (2013), How firm managers understand innovation: implications for the design of innovation surveys, in F. Gault (ed.), *Handbook of Innovation Indicators and Measurement*, Edward Elgar Publishing, 88–108.

Ashouri, S., A. Suominen, A. Hajikhani, L. Pukelis, T. Schubert, S. Türkeli, C. Van Beers, S. Cunningham (2021), *Indicators on Firm Level Innovation Activities from Web Scraped Data*, https://doi.org/10.34894/W3W2JQ

Axenbeck, J., P. Breithaupt (2021), Innovation indicators based on firm websites: which website characteristics predict firm-level innovation activity? *PLoS One* 16(4), https://doi.org/10.1371/journal.pone.0249583.

Back, B., J. Toivonen, H. Vanharanta, A. Visa (2001), Comparing numerical data and text information from annual reports using self-organizing maps, *International Journal of Accounting Information Systems* 2(4), 249–269.

Banerji, D., T. Reimer (2019), Startup founders and their LinkedIn connections: are well-connected entrepreneurs more successful? *Computers in Human Behavior* 90, 46–52.

Bhimani, H., A.L. Mention, P.J. Barlatier (2019), Social media and innovation: A systematic literature review and future research directions, *Technological Forecasting and Social Change* 144, 251–269.

Blazquez, D., J. Domenech, A.M. Debón Aucejo (2018), Do corporate websites' changes reflect firms' survival? *Online Information Review* 42(6), 956–970.

Breithaupt, P., R. Kesler, T. Niebel, C. Rammer (2020), *Intangible Capital Indicators Based on Web Scraping of Social Media*, ZEW Discussion Paper No. 20-046, Centre for European Economic Research.

Bruhn, M., V. Schoenmueller, D.B. Schäfer (2012), Are social media replacing traditional media in terms of brand equity creation? *Management Research Review* 35(9), 770–790.

Castellacci, F., J.M. Natera (2012), Innovation surveys in Latin America: a primer, Inno*vation and Development* 2(1), 199–204.

Chiang, J.K.-H., H.-Y. Suen (2015), Self-presentation and hiring recommendations in online communities: lessons from LinkedIn, *Computers in Human Behavior* 48, 516–524.

Chung, S., A. Animesh, K. Han, A. Pinsonneault (2015), The Business value of firms' social media efforts: evidence from Facebook, *Proceedings of the 17th International Conference on Electronic Commerce* 2015, 1–8.

Cirera, X., S. Muzi, (2020). Measuring innovation using firm-level surveys: evidence from developing countries, *Research Policy* 49(3), 103912.

Cohen, W.M. (2010), Fifty years of empirical studies of innovative activity and performance, in B.H. Hall, N. Rosenberg (eds.), *Handbook of the Economics of Innovation, Volume 1*, Elsevier, 129–213.

Cojoianu, T.F., G.L. Clark, A.G. Hoepner, P. Veneri, D. Wójcik (2020), Entrepreneurs for a low carbon world: how environmental knowledge and policy shape the creation and financing of green start-ups, *Research Policy* 49(6), 103988.

Coombs, R., P. Narandren, A. Richards (1996), A literature-based innovation output indicator, *Research Policy* 25(3), 403–413.

Coursaris, C.K., W. van Osch, B.A. Balogh (2016), Do Facebook likes lead to shares or sales? Exploring the empirical links between social media content, brand equity, purchase intention, and engagement, *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS),* 3546–3555.

Crass, D. (2014a), *The Impact of Brand Use on Innovation Performance – Empirical Results for Germany*, ZEW Discussion Paper No. 14-119, Centre for European Economic Research.

Crass, D. (2014b), *Which Firms Use Trademarks – and Why? Representative Firm-Level Evidence from Germany*, ZEW Discussion Paper No. 14-118, Centre for European Economic Research.

Daas, P.J.H., S. van der Doef (2021), *Using Website texts to detect Innovative Companies*, Center for Big Data Statistics Working Paper No. 01-21, Statistics Netherlands.

Daas, P.J.H., S. van der Doef (2021), Detecting innovative companies via their website, *Statistical Journal of the IAOS* 36(4), 1239–1251.

Dalle, J., M. Besten, C. Menon (2017), *Using Crunchbase for Economic and Managerial Research*, OECD Science, Technology and Industry Working Papers, No. 2017/08, OECD Publishing.

Gandin, I., C. Cozza (2019), Can we predict firms' innovativeness? The identification of innovation performers in an Italian region through a supervised learning approach, *PLoS One* 14(6), 1–16.

Garechana, G., R. Río-Belver, I. Bildosola, M. Rodríguez Salvador (2017), Effects of innovation management system standardization on firms: evidence from text mining annual reports, *Scientometrics* 111, 1987–1999.

Geroski, P.A., J. Van Reenen, C.F. Walters (1997), How persistently do firms innovate? *Research Policy* 26(1), 33–48.

Gök, A., A. Waterworth, P. Shapira (2015), Use of web mining in studying innovation, *Scientometrics* 102(1), 653–671.

Gortmaker, J., J. Jeffers, M. Lee (2021), *Labor Reactions to Credit Deterioration: Evidence from LinkedIn Activity*, available at SSRN (http://dx.doi.org/10.2139/ssrn.3456285).

Griliches, Z. (ed.) (1984), R*&D, Patents, and Productivity*, University of Chicago Press.

Griliches, Z. (2007), Patent statistics as economic indicators: a survey, in Z. Griliches (ed.), *R&D and Productivity: The Econometric Evidence*, University of Chicago Press, 287–343.

Hamilton, R.H., H.K. Davison (2018), The search for skills: Knowledge stars and innovation in the hiring process, *Business Horizons* 61(3), 409–419.

Hong, S., L. Oxley, P. McCann (2012), A survey of the innovation surveys, *Journal of Economic Surveys* 26(3), 420–444.

Ikeuchi K. (2017), Measuring innovation in firms, in Y. Honjo (ed.), *Competition, Innovation, and Growth in Japan*, Springer, 77–97.

Ji, Y., O. Rozenbaum, K. Welch (2017), *Corporate Culture and Financial Reporting Risk: Looking Through the Glassdoor*, available at SSRN (doi: 10.2139/ssrn.2945745).

Kahn, K.B. (2018), Understanding innovation, *Business Horizons* 61(3), 453–460.

Kemeny, T., M. Nathan, B. Almeer (2017), *Using Crunchbase to Explore Innovative Ecosystems in the US and UK*, Discussion Paper Series, University of Birmingham.

Kim, J.H. (2012), A hyperlink and semantic network analysis of the triple helix (university-government-industry): the interorganizational communication structure of nanotechnology, *Journal of Computer-Mediated Communication* 17(2), 152–170.

Kinne, J., J. Axenbeck (2018), *Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany*, ZEW Discussion Paper No. 18-033, Centre for European Economic Research.

Kinne, J., J. Axenbeck (2020), Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study, *Scientometrics* 125, 2011–2041.

Kinne, J., D. Lenz (2019), *Predicting Innovative Firms Using Web Mining and Deep Learning*, ZEW Discussion Paper No. 19-001, Centre for European Economic Research.

Kinne, J., D. Lenz (2021), Predicting innovative firms using web mining and deep learning, *PLoS One* 16(4), https://doi.org/10.1371/journal.pone.0249071.

Kinne, J., B. Resch (2018), Generating big spatial data on firm innovation activity from text-mined firm websites, *GI_Forum* 1, 82–89.

Kinne, J., M. Krüger, D. Lenz, G. Licht, P. Winker (2020), *Coronavirus Pandemic Affects Companies Differently. A High-Frequency Website Analysis of Companies' Reactions to the Coronavirus Pandemic in Germany*, ZEW Expert Brief 20-05, Centre for European Economic Research.

Kleinknecht, A., H.J. Reinders (2012), How good are patents as innovation indicators? Evidence from German CIS data, in M. Andersson, B. Johansson, C. Karlsson, H. Lööf (eds.), *Innovation and Growth: From R&D Strategies of Innovating Firms to Economy-Wide Technological Change*, Oxford University Press, 115–127.

Kleinknecht, A., J.O. Reijnen, W. Smits (1993), Collecting literature-based innovation output indicators. The experience in the Netherlands, in D. Bain, A. Kleinknecht (eds.), *New Concepts in Innovation Output Measurement*, Palgrave Macmillan, 42–84.

Kleinknecht, A., K. van Montfort, E. Brouwer (2002), The non-trivial choice between innovation indicators, *Economics of Innovation and New Technology* 11(2), 109–121.

Klinger, J., J. Mateos-Garcia, C. Stathoulopoulos, C. Tippett, R. Moeremans, J. Morret (2018), *Exploratory Report B: Toward the Incorporation of Big Data in the European Innovation Scoreboard*. European Innovation Scoreboard project 2018-2019, Brussels.

Krieger, B., M. Pellens, K. Blind, S. Gruber, T. Schubert (2021), Are firms withdrawing from basic research? An analysis of firm-level publication behaviour in Germany, *Scientometrics* 126, 9677–9698.

Krüger, M., J. Kinne, D. Lenz, B. Resch (2020), *The Digital Layer: How Innovative Firms Relate on the Web*, ZEW Discussion Papers No. 20-003, Centre for European Economic Research.

Li, Y., S. Arora, J. Youtie, P. Shapira (2018), Using web mining to explore Triple Helix influences on growth in small and mid-size firms, *Technovation* 76/77, 3–14.

Libaers, D., D. Hicks, A.L. Porter (2010), A taxonomy of small firm technology commercialization, *Industrial and Corporate Change* 25(3), 371–405.

Mendonça, S., T.S. Pereira, M.M. Godinho (2004), Trademarks as an indicator of innovation and industrial change, *Research Policy* 33(9), 1385–1404.

Meyer-Krahmer, F. (1984), Recent results in measuring innovation output, *Research Policy* 13, 175–182.

Mirończuk, M., J. Protasiewicz (2016), A diversified classification committee for recognition of innovative internet domains, in S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, D. Kostrzewa (eds.), *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* (Communication in Computer and Information Sciences 613), Springer, 368–383.

Mirtsch, M., J. Kinne, K. Blind (2020), Exploring the adoption of the international information security management system standard iso/iec 27001: a web mining-based analysis, *IEEE Transactions on Engineering Management* 68(1), 87–100.

Misirlis, N., M. Vlachopoulou (2018), Social media metrics and analytics in marketing – S3M: a mapping literature review, *International Journal of Information Management* 38(1), 270–276.

Mortenson, P.S. (2008), *The Regionalisation of CIS Indicators: The Danish Experience*, paper presented at the 32nd CEIES Seminar, 'Innovation indicators–more than technology', Aarhus.

Nakatsuji, M., Y. Miyoshi, Y. Otsuka (2006), Innovation detection based on user-interest ontology of blog community, *International Semantic Web Conference (ISWC2006)*, 515–528.

Nakatsuji, M., M. Yoshida, T. Ishida (2009), Detecting innovative topics based on user interest ontology, *Web Semantics: Science, Services and Agents on the World Wide Web* 7(2), 107–120.

Nelhans, G. (2020), *Analysis of Company, EU Projects, Policy Documents, Clinical Guideline and Social Media/Media Data Report*, Deliverable 5.2 of the project "Big Data approaches for improved monitoring of research and innovation performance and assessment of the societal impact in the Health, Demographic Change and Wellbeing Societal Challenge" (Data4Impact), available at https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5d1e1ff02&appId=PPGMS.

Niebel, T., F. Rasel, S. Viete (2019), BIG Data - BIG gains? Understanding the link between big data analytics and innovation, *Economics of Innovation and New Technology* 28(3), 296–316.

Obschonka, M., D.B. Audretsch (2020), Artificial intelligence and big data in entrepreneurship: a new era has begun, *Small Business Economics* 55, 529–539.

Pavitt, K. (1985), Patent statistics as indicators of innovative activities: possibilities and problems, *Scientometrics* 7(1-2), 77–99.

Pisano, S., L. Lepore, R. Lamboglia (2017), Corporate disclosure of human capital via LinkedIn and ownership structure, *Journal of Intellectual Capital* 18(1), 102–127.

Pukelis, L., V. Stanciauskas (2019), *Using Internet Data to Compliment Traditional Innovation Indicators*, paper presented at the International Conference on Public Policy (ICPP4), available at https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf.

Rammer, C., D. Czarnitzki, A. Spielkamp (2009), Innovation success of non-R&D-performers: substituting technology by management in SMEs, *Small Business Economics* 33(1), 35–58.

Rietsch, C., C. Beaudry, M. Héroux-Vaillancourt (2016), Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community, *First International Conference on Advanced Research Methods and Analytics (CARMA2016)*, Valencia, 100–115.

Romanov, D., M. Ponfilenok, N. Kazantsev (2013), Potential innovations (new ideas/trends) detection in information network, *International Journal of Future Computer and Communication* 2(1), 63–66.

Ruhrmann, H., M. Fritsch, L. Leydesdorff (2021), Synergy and policy-making in German innovation systems: smart Specialisation Strategies at national, regional, local levels? *Regional Studies* (doi: 10.1080/00343404.2021.1872780).

Schmoch, U. (2003), Service marks as novel innovation indicator, *Research Evaluation* 12(2), 149–156.

Schmoch, U., S. Gauch (2009), Service marks as indicators for innovation in knowledge-based services, *Research Evaluation* 18(4), 323–335.

Schubert, T., A. Jäger, S. Turkeli, F. Visentin (2020), *Addressing the Productivity Paradox with Big Data: A Literature Review and Adaptation of the CDM Econometric Model*, UNU-MERIT Working Paper 2020-050, Maastricht University.

Sloan, L., A. Quan-Haase (eds.) (2017), *The SAGE Handbook of Social Media Research Methods*, Sage.

Som, O. (2012), *Innovation Without R&D: Heterogeneous Innovation Patterns of Non-R&D-performing Firms in the German Manufacturing Industry*, Springer Science & Business Media.

Suominen, A., A. Hajikhani (2021), Research themes in big data analytics for policymaking: insights from a mixed-methods systematic literature review, *Policy Internet* 2021, 1–21.

Tether, B.S. (2002), Who co-operates for innovation, and why: an empirical analysis, *Research Policy* 31(6), 947–967.

Tirunillai, S., G.J. Tellis (2012), Does online chatter really matter? Dynamics of user-generated content and stock performance, *Marketing Science* 31(2), 198–215.

Veltri, G.A. (2013), Microblogging and nanotweets: nanotechnology on twitter, *Public Understanding of Science* 22(7), 832–849.

Von Bloh, J., T. Broekel, B. Özgun, R. Sternberg (2020), New(s) data for entrepreneurship research? An innovative approach to use Big Data on media coverage, *Small Business Economics* 55, 673–694.

Youtie, J., D. Hicks, P. Shapira, T. Horsley (2012), Pathways from discovery to commercialisation: using web sources to track small and medium-sized firm strategies in emerging nanotechnologies, *Technology Analysis & Strategic Management* 24(10), 981–995.

Yu, Y., L. Parillo-Chapman, M. Moore (2020), *Fashion Printing Technology Diffusion: Big Data Analytics*, Internationale Textile and Apparel Association Annual Conference Proceedings (Vol. 77, No. 1), Iowa State University Digital Press.

Zide, J., B. Elman, C. Shahani-Denning (2014), LinkedIn and recruitment: how profiles differ across occupations, *Employee Relations* 36(5), 583–604.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

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