Using Internet Data to Compliment Traditional Innovation Indicators

Lukas Pukelis[1] and Vilius Stanciauskas[2]

[1]*lukas.pukelis@ppmi.lt*
Public Policy and Management Institute (PPMI), Gedimino g. 50, 01110, Vilnius (Lithuania)

[2]*vilius@ppmi.lt*
Public Policy and Management Institute (PPMI), Gedimino g. 50, 01110, Vilnius (Lithuania)

Abstract
Currently, innovation activities of enterprises are assessed using either surveys or IPR/patent analysis. Neither of these methods are ideal for a variety of reasons, including high cost, high labour intensity, and certain problems with data reliability. This paper presents our efforts to develop new innovation indicators using internet data and supervised machine learning, to be used alongside the existing measures. The paper draws on the insights of the "Data4Impact" project, which sought to explore the possibilities of using Big Data methodologies to assess the impact of research funding. Initial results suggest that the method described in here, though not without its shortcomings, shows certain promise to capture innovation activities of enterprises. Its main advantages are the ability to capture non-patented innovations and to do so in a fraction of time compared to traditional survey approaches. This might benefit research funding bodies which seek to broaden their impact monitoring measures.

Introduction
Being able to assess and estimate the innovation activities of enterprises in the private sector is a highly salient question for scholars and policy-makers. Innovation is directly related to economic growth (Rosenberg 2004) and as such is one of the desired effects of the research-funding or certain economic policy measures. However, the tools available to scholars to measure the innovation among the private enterprises are limited and not ideal. Currently, it is mostly done by using innovation surveys or intellectual property (patent) analysis. Each of these methods when used in isolation or together can present a rich and detailed picture of innovation activities of private enterprises. However, they have several serious shortcomings that can impede their use. The main weakness of patent analysis is that patents do not directly correspond to innovations in a sense that: not all patented ideas become innovations; not all innovations are patented; propensity to patent differs among enterprises, and patents in different jurisdictions are not directly comparable (Archibugi & Planta 1996). Meanwhile, the main weakness of using surveys is the cost in terms of time and labour resources needed to carry them out. Currently, the most respected effort to carry out a wide-scale innovation survey in Europe is the "Community Innovation Survey" (CIS) by Eurostat. The main shortcoming of CIS data is that data collection, processing and publication can take up to four years (e.g. in 2018 most recent CIS data was from 2014). Even with smaller surveys the time lag between the initiation of the survey to the data can be considerable, often taking months and requiring hundreds of man-hours of labour. Particularly worrying recent trend is the declining survey response rates and survey fatigue, especially among the organisations benefitting from the EU research funding.

Given these shortcomings of existing methodologies for estimating innovations, there has been some recent interest of using internet data and other Big Data methodologies to derive additional indicators to estimate innovations or other impacts of research funding. For instance, Centre for European

Economic Research in Manheim University has explored using internet data to track the innovation activities of German enterprises (Kinne & Axenbeck 2018) and the European Commission has launched several initiatives to use Big Data methodologies to estimate the impact of research funding (EC 2015).

This paper presents partial results of one such project – "Data4Impact" funded under H2020 Co-Creation programme. While the project's overall goal was to use Big Data methods to provide a comprehensive summary of the outputs, results, and potential impacts of EU-funded projects, this paper concentrates on a single indicator – using internet data to estimate the number innovations produced by participating companies. The paper is structured as follows: the first part presents the overall rationale why company websites could be considered a valid data source and presents the pros and cons on using the internet data. The second part outlines the methodology for identifying innovation content in company websites and the key challenges of working with internet data. The third part briefly describes the results and benchmarks the company innovation counts from the internet to the 'hard' patent data.

Internet as a data source

There have been several attempts to use internet data to estimate innovations among companies in the private sector (Kinne & Axenbeck 2018). Nonetheless, the internet data is still considered novel and as such it poses certain validity and reliability concerns. Namely, there are two main questions regarding internet data or data from company websites: the *validity question* – do the indicators obtained from company websites correspond to the number of 'true' innovations; and *reliability* question – can we estimate the number of innovations in a reliable manner. The reliability question is purely technical in nature and is discussed in greater detail in the subsequent sections of the paper. Meanwhile, the validity question is more philosophical and can hardly be answered in a simple and straight-forward manner. Instead, we present an argument that the overall validity of the internet data is just as valid as the survey data.

First, over 85% of all enterprises in the EU have a website or some presence on the web[1]. Though the coverage is not universal, for all practical purposes it is large enough that we could consider that data on a certain enterprise could in principle be found and accessed on the web. Furthermore, there are no obvious biases in terms of countries, regions or economic sectors that could jeopardise the validity of the data.

Second, enterprises increasingly view web as an important platform to supply information about themselves and their activities. There are already some studies carried out that demonstrate that enterprises do post information on their innovation activities on the web and that overall it is broadly comparable to the innovation data from other sources (Gök, Waterworth & Shapira 2015; Katz & Cothey 2006; Youtie et al 2012; Aurora et al. 2013). Having said that, it is important to point out that there are certain peculiarities associated with using data from the company websites. It is important to note that the data on company websites is presented there to communicate the essential information about the enterprise in question to its target audiences, which might include: clients, business partners, and/or competitors. Scientists and policy analysts are not the target audience and the information on the websites is not tailored to their needs. This manifests in a variety of ways, for instance companies tend to focus on their products that are available for sale at the moment or will be in the immediate future. R&D activities with less immediate market applications tend to receive less coverage (Gök, Waterworth & Shapira 2015).

[1] Eurostat: Digital economy and society statistics – enterprises <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_enterprises>

Furthermore, another significant issue arises from the fact that company website data is self-reported data. This means that different companies might have different propensity to announce their innovation activities or might apply different threshold to what exactly defines innovation (compared to e.g. incremental improvement). However, though definitions of what constitutes a genuine innovation might differ, enterprises do have a market incentive not to understate or overstate their innovation activities. Since companies with a propensity to under or over-sell their position would face negative reactions from partners and/consumers, enterprises are under certain pressure to provide accurate information on their webpages.

Though self-reporting this is a serious shortcoming, it is by no means unique to the internet data. Innovation surveys also rely on self-reporting and thus suffer from the same validity concerns. It is important to note that in surveys concerns over self-reporting are usually addressed by increasing the sample sizes. It is hoped that in a large sample different biases will cancel each other out and the indicators would be generally valid. However, in surveys ramping up the samples is often difficult and costly, whereas with internet data these marginal costs are practically non-existent and data collection from the web can be carried out for a fraction of the cost, while maintaining either larger samples or eliminating them altogether and carrying out full population studies.

Table 1 contains a summary of pros and cons of using different information sources to measure enterprise innovation. Compared to other sources, internet data has the advantage of being relatively cheap and rapid. This means that indicators computed using these data can be 'refreshed' more frequently and without major costs to either data collectors or enterprises themselves. However, that comes at a cost of lower granularity. As outlined in the next section, inferring innovation counts from company websites involves extensive natural language processing and working with free-text data. There is an enormous amount of variation in how companies structure their websites and content within. As such identifying innovation related content on company websites is a hard-enough task and classifying innovation content into smaller categories becomes increasingly complicated.

Table 1. Pros and cons of different sources to measure enterprise innovation

Method	*Pro*	*Con*
Survey	• Ability to directly ask desired questions directly to the target respondent; • Granularity – being able to go into detail	• Resource intensive; • Self-report bias; • Survey fatigue and declining response rates; • Recency bias – can only inquire about recent occurrences; • Low technological detail.
Patent/IPR analysis	• 'Hard data' – no self-reporting bias; • Ability to go back in time decades or centuries; • High technological detail.	• Fuzzy link between patents and innovations – not all innovations are patented and not all patents are innovations.
Internet data	• Cheap and rapid; • Possibility to have large samples/ full population studies;	• Lower granularity; • Reusing data originally intended for different

	• Being able to go back in time months/years.	purpose and different audience.

Methodology

Defining and operationalising the concept of innovation for internet data

Oslo Manual considers innovation to be a continuous process, which can manifest through a variety of different types as product, process, marketing or organisational innovations (Oslo Manual 2018). Following the taxonomy of innovation activities of the Oslo Manual, the most reputable source of company innovation data in the EU – Community Innovation Survey also distinguishes between a variety of different innovation activity types. While, it is possible to collect data on the different types of innovation activities in the desired level of detail using the survey approach, it is not the case with internet data. Gathering data from the web, means that texts originally composed for different purposes are used to collect data on and measure company innovations. This in turn means that it is simply not possible to extract the same level of detail from these texts as is possible with the survey approach. As companies are entities with a clear goal to sell their products and services, they dedicate the biggest portion of their websites for describing them. Similarly, they can expose the user to the end product a marketing campaign without revealing any details on whether the campaign itself contained any marketing innovations.

Furthermore, as mentioned in the previous section, in their websites, companies tend to focus on innovation outputs – new products, services and improvements in the process rather then innovation activities with no immediate market application because these directly contribute to their core business. As such, in our methodology, we focussed solely on the innovation outputs and particularly on the "innovation announcement texts". We define "innovation announcement texts" as any massage in the company web domain that explicitly states that a company introduces a new product/service or improvement in the internal processes. By doing this, we are restricting our focus to only very explicit innovation announcements and risk not identifying such innovations that happened but were not explicitly presented as innovations. However, we consider this restriction necessary to ensure the providence and accuracy of our indicator. A few sample innovation announcement texts are presented in Table 2.

Table 2. Sample innovation announcement texts

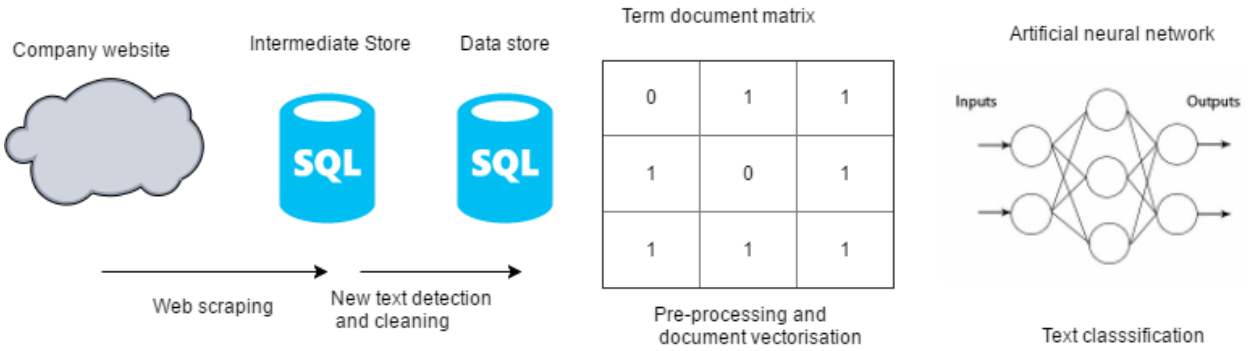
CENTUM® VP R6.05 Integrated Production Control System - With a new processor module and an enhanced engineering function - CENTUM VP R6.05 Integrated Production Control System Yokogawa Electric Corporation (Tokyo: 6841) announces that it will release CENTUM® VP R6.05, an enhanced version its flagship integrated production control system, on October 24.	Systematic introduces new capabilities to SitaWare and IRIS solutions The new functions include 3D visualisation and will improve situational awareness, safety, and usability, among other benefits
Unigraf is introducing UCD-340, world's first integrated test equipment for testing DisplayPort™	Datalogic a g lobal leader in the automatic data capture and process automation markets, proudly announces release of IMPACT Software 12.0, the latest version of the well-known software by Datalogic for Vision Guided Robotics applications.

In other words, product (or service) innovations tend to leave an observable and detectable trace on company websites, whereas many other types of innovations do not. This is a core feature of using internet data to measure innovations: the measurement is constrained to a segment of innovation activities – innovation outputs which are described and presented on the websites. In our measurement, we used a 2x2 matrix to differentiate between different kinds of innovation outputs. On one axis we distinguished between those products that are already on the market and can be bought or sold and those which are not on the market yet, such as products with scheduled launch date, drugs currently undergoing clinical trials or prototype/demonstrator versions of possible new products. On the other axis we differentiated between products that are tangible (have such physical characteristics as volume and mass) and those which are intangible – having no observable volume or mass. Products such as devices, tools, or consumables would be classified as tangibles, while software, databases or various services would fall into intangible category. This distinction is outlined in the innovation-output matrix below.

Data gathering and analysis

The methodology for gathering and analysing the data is presented in Figure 1. In essence, this method utilises supervised machine learning to recognise innovation related texts from company website and label them as such. For this method to work, it first requires a pre-labelled sample of innovation and non-innovation texts on which the machine learning model is trained. Prior to deploying the model, we have manually labelled texts from 500 enterprises and trained the model on the labelled sub-set. The model was trained to distinguish between two categories: innovation related text and non-innovation related items.

Figure 1. Workflow for harvesting and analysing internet data



Company innovation indicators are obtained in six steps:

1. Company websites are scraped and all textual data wherein is collected;
2. Data is placed in intermediate storage and new data is identified;
3. New data is indexed and placed in main data store;
4. Texts are pre-processed and vectorised;
5. Artificial network model is used to classify each text as innovation relate or not;
6. Company innovation scores are computed from aggregate data.

In the first step we crawled and scraped the company websites and collected text data wherein. To comply with the data protection regulation, we *ex ante* specified sections of websites which were ignored by the scraper, such as pages for contact information, biographies of company employees,

etc. Where available instructions for scrapers located in *robots.txt* were obeyed. If an enterprise employed any kind of scraper blocks, they were obeyed and no measures were taken to counter them. Overall, we successfully scraped the websites of 1301 companies of oou1392 company sample, resulting in the response rate of 93,5%.

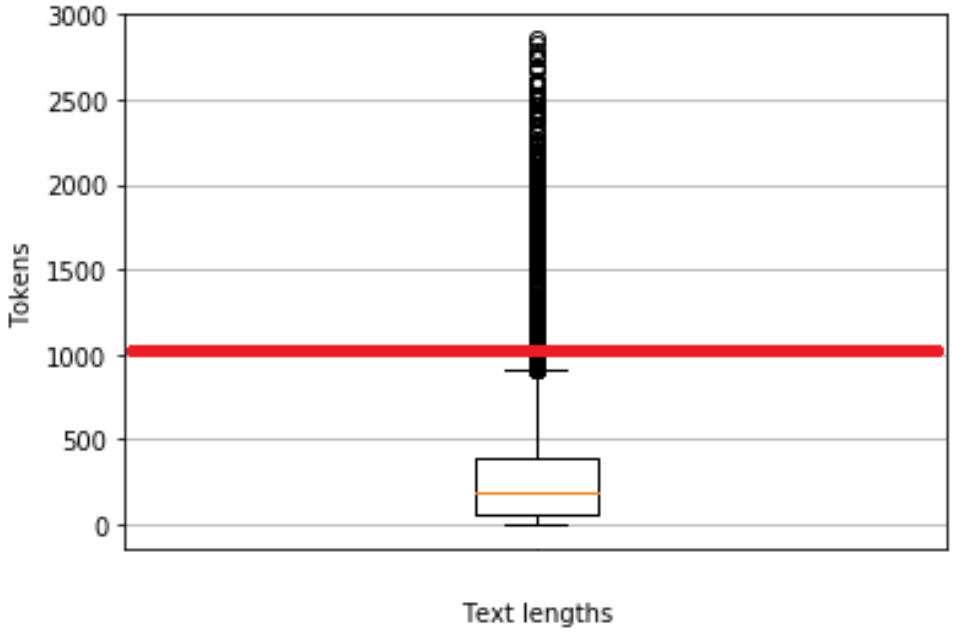
Scraped texts were placed in the intermediate data store. From there they were compared to the existing records (from previous scraping iterations) and new or partially new content was identified and placed for further analysis. Additionally, if some of the existing records were not found in the current iteration, they were marked as discontinued in the main database.
After the standard text pre-processing procedure (tokenization, lemmatization, stop-word and punctuation removal) texts were vectorized, i.e. converted to document term matrixes using '*doc2idx*' approach, which replaces text word tokens with corresponding numeric values, see Table 3.

Table 3. Illustration of doc2idx approach

Normal text	"A fool thinks himself to be wise, but a wise man knows himself to be a fool."
Tokens	'fool', 'thinks', 'be', 'wise', 'wise', 'man', 'knows', 'be', 'fool'
Dictionary	'fool': 1, 'thinks': 2, 'be': 3, 'wise': 4, 'man': 5, 'knows': 6
Vector	1,2,3,4,5,6,3,1

In this step we limited the size of dictionary to 50 000 entries. In developing the dictionary, we filtered out the extreme values: words that either appeared fewer than four times or words that appeared in more than 50% of the documents. After filtering the extremes, we selected 50 000 most common items. We arrived at 50 000 mark iteratively: our experiments with the labelled data suggested that the dictionary of this size is sufficient to get decent performance from the model, while ensuring that the model does not become unnecessarily large. We also standardized the texts to the fixed length of 1000 tokens. Texts shorter than the mark were padded with zero values, while longer texts were truncated, leaving first 1000 tokens. Similarly, 1000 token length was decided after reviewing the structural characteristics of the labelled sample, which revealed that 1000 token mark roughly corresponds to 95^{th} percentile of the document length distribution (see Figure 2).

Figure 2. Distribution of texts by length



Once the document is vectorised, it can be put into the artificial neural network (ANN) model to be classified. These models take in vectorised documents and output a probability that a particular text is innovation-related. That way it is possible to take into consideration for how the word meanings might change on account of its context.

Overall, we chose the artificial neural network as the main model for text classification because these models can solve highly non-linear problems and, therefore, are almost uniquely suited to perform free-text classification tasks, where the complexity of data is extremely high. Prior to settling down on the text classification algorithm, we performed extensive testing, comparing the performance of the most commonly used classification algorithms (see Table 4).

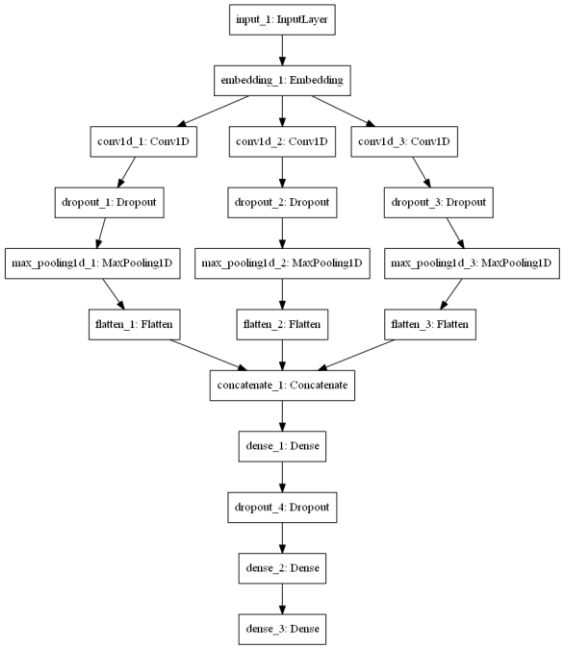
Table 4. Comparison of different classification algorithms on the labelled data sample

N: 163 Companies; 31 898 webpages Training set: 23 923 (75%); Test set: 7 975 (25%)		*Logistic regression*		*Random forest*		*Support Vector Machine*		*ANN*	
		Model prediction		Model prediction		Model prediction		Model prediction	
		Neg.	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.	Pos.
Actual	Neg.	7397	52	7403	46	7448	1	7442	7
Value	Pos.	166	360	159	367	448	78	41	485
F1 score:		0.767		0.781		0.257		0.952	

As indicated in Table 4, two other models: Logistic regression and Random Forrest also performed comparatively well, but their predictions were on average 20% less accurate than those of the ANN model. Most importantly, these models generated significant numbers of false negatives (model fails to identify innovation content), which is particularly problematic because, if need be, false positive predictions can be addressed and filtered out in later stages of the analysis, while false negatives mean that some innovation content remains permanently uncaptured.

We used a convolutional neural network (LeCun & Bengio 1995) because it evaluates not individual words, but phrases of specified length. More specifically, we constructed a convolutional neural network for text classification similar to the one describer by Kim (2014). The ANN model has a single dynamic embedding layer, followed by three parallel convolutional-dropout-pooling clusters with varying window sizes. These are followed by a concatation layer and two dense layers with a dropout layer in between. The overall schematic for the ANN model used is depicted in Figure 3.

Figure 3. Schematic for the ANN model



After using the model to generate predictions for the whole dataset, we performed additional validation tests for model performance. A sample of 1 000 model predictions with equal balance between classes was drawn from the dataset these texts were manually checked and manual check assessment was compared with the model prediction. We discovered five cases of false positives and none false negatives. It was judged that the model performance is satisfactory and the model design was frozen with the same configuration in place.

Results

Share of Innovative companies

Our measurement yields that 588 companies out of 1301 participants have produced at least one innovation resulting in innovation rate of 45%, see Table 5. Though this figure is slightly smaller than the 'Share of Innovative Companies in EU-28' from CIS survey, the two figures are not directly comparable, as CIS utilises a much broader definition of innovation. If we look at the share of product innovative enterprises in CIS the figure is much smaller – just below 24% percent. Overall, out results are consistent with the expectations that companies benefiting from EU research funding (as were the target group of the Data4Impact project) would have a higher share of innovating enterprises than the general enterprise population.

Table 5. Share of innovative companies and innovation content

	Records in DB	*Companies*	*CIS innovative companies*	*CIS product innovative companies*
All	567554	1301		
With Innovation	13815	588		
Share	2.5%	45%	49%	23.9%

Source: Own calculations; CIS 2014 data

Innovations from internet and patents

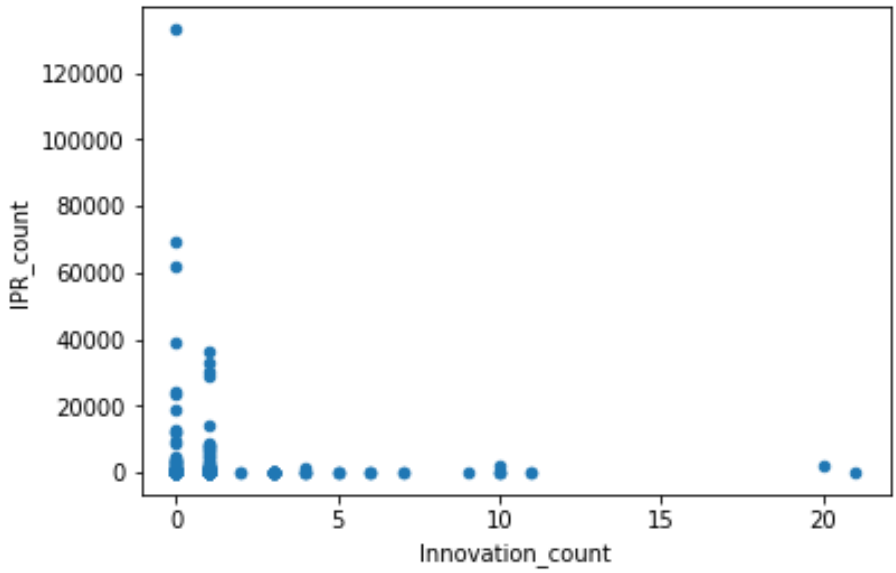
We compared our innovation output scores to other commonly used measures for company innovation, more specifically company patent counts. We compared the number of distinct company innovations to the overall number of patents and patent applications filed by the company. As shown in Figure 4 and Table 6, there is almost no correlation between the patent and web data.

Table 6. Correlation between Innovation Outputs and IPR

Pearson's r	0.0713

Source: Lens.org, own calculations

Figure 4. Innovation Outputs and IPR



Our findings closely echo previous studies on the topic, which also noted poor correspondence between patent and internet data (Gök, Waterworth, Shapira 2015). In addition to the aforementioned reasons why patents might not directly correspond to innovations (different motivations or propensity to patent, etc.), we must note two additional factors:

1. A significant number of innovations have a significant service component, which does not relate to patents in any way. For instance, an internet service provider might introduce an innovative service (e.g. 5G internet connection) without owning a single 5G-related patent.
2. In large multinational corporations a dedicated research division, which might have a significant number of patents might not have its own dedicated website (e.g. research arm of the Microsoft Corporation does not have its own website, only a subdomain in www.microsoft.com).

Overall, these findings show that internet data captures different aspects of innovation activities than the IPR indicators and, hence, can be used to complement them in future research.

Participant-level results
In the "Data4Impact" project, we analysed the population of companies, which took part in health-themed EU research projects in FP7 and H2020. Since large pharmaceutical companies are frequent participants of these projects, it is not surprising that the highest numbers of innovation announcement texts are found in the websites of these companies as indicated in Table 7. The results are in-line with the initial expectations and the EC monitoring data. For a more detailed look, please visit the Data4Impact project website.[2]

Table 7. Top-3 companies with the most innovation announcement texts

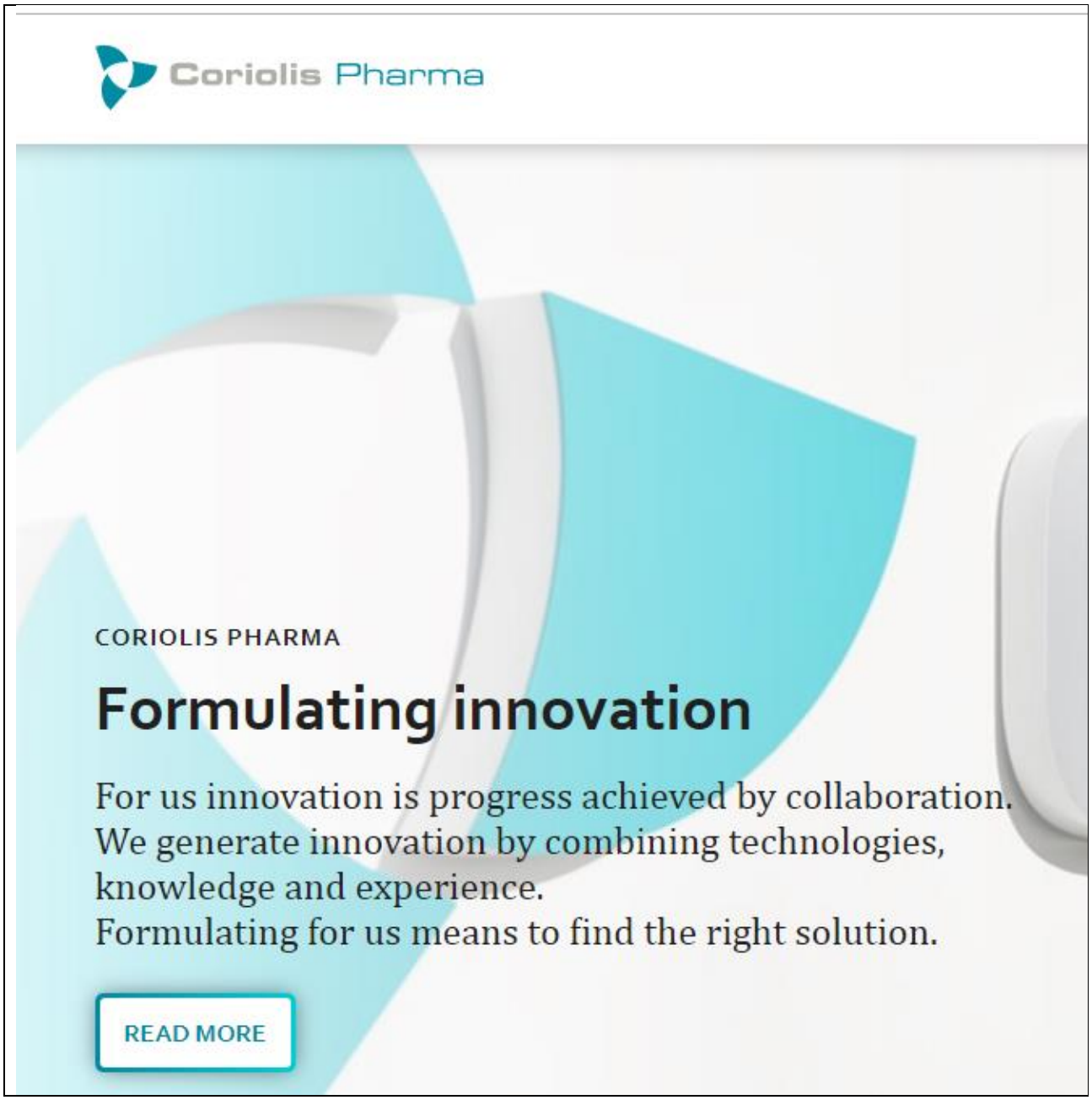
Company	**Innovation texts**
Astra Zeneca	2 153
Agilent	364
Gilead	330
Oxford BioMedica	300
Merck	298
GSK	298
Unilever	293
Novartis	159

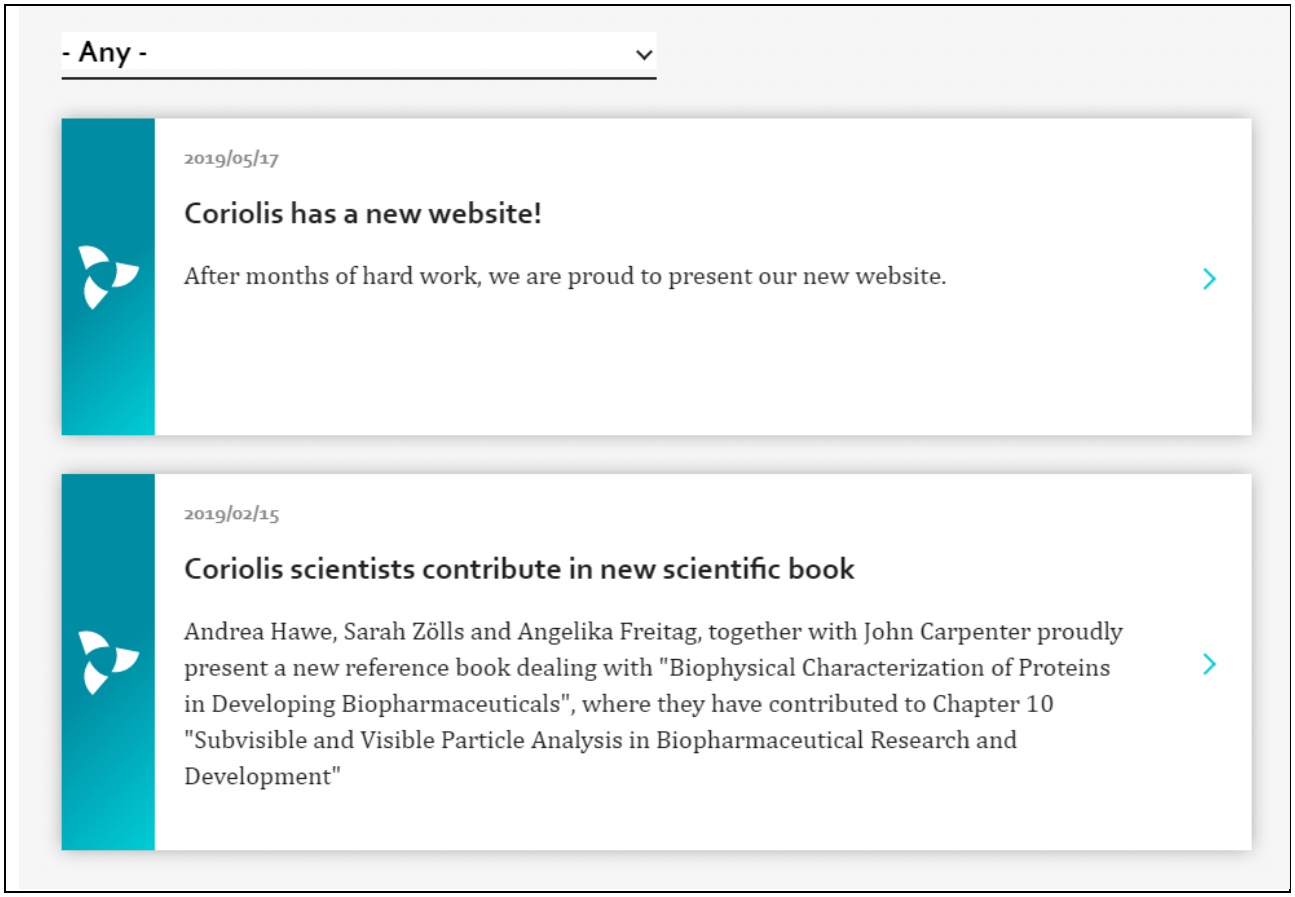
It is worth noting that though our approach has detected the most innovation announcement texts in the website of Astra Zeneca, it cannot be definitively stated that AZ is significantly more innovative than the other companies in the table. The biggest reason why the indicator value for AZ is so high is that the AZ website has a long and thorough news archive which stretches back for about a decade, while other websites, e.g. that of Novartis, does not have such features. In its current state, the vast indicator value difference between AZ and other companies may not reflect the true innovation performance of the companies. However, this can be addressed by carrying out repeated data collection and analysis rounds on the company panel and measuring not the absolute number of innovation texts found, but the rate at which new innovation announcements appear on the website.

Our focus on very explicit innovation output texts meant that we are better able to capture companies which focus on delivering products to the market (rather than those carrying out early or mid-stage research) and those which put out more news announcements related to their activities. For instance, using our approach we did not find a single innovation output text in the website of "Coriolis Pharma" (https://www.coriolis-pharma.com/), even though the company has the word "innovation" in their slogan. However, a closer look reveals that their news feed is empty and the company releases very few news announcements, see Figure 5.

[2] Data4Imnpact project website < http://www.data4impact.eu/>

Figure 5. Case Study "Coriolis Pharma"





Summary

This paper presented some of the results of the 'Data4Impact' project, aimed at utilising Big Data methodologies to estimate the impact of research funding. It summarised our experience in using internet as a source of company innovation data.

Overall, we consider that internet data, especially data from company websites, can be treated as a valuable source of company innovation data. Though further improvements to the data gathering and analysis methodology are needed, preliminary results indicate that mentions of innovation outputs can be captured from the company websites with high degree of accuracy using supervised machine learning techniques. However, as this is one of the first attempts to leverage internet data to capture company innovation, it naturally needs more fine-tuning and refinement before it can reach its full potential.

Funding acknowledgements:

This work is based on research funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531.

References:

Rosenberg, N., 2004. Innovation and economic growth. Innovation and Economic Growth, 52.

Archibugi, D. and Planta, M., 1996. Measuring technological change through patents and innovation surveys. Technovation, 16(9), pp.451-519.

Kinne, J. and Axenbeck, J., 2018. Web mining of firm websites: A framework for web scraping and a pilot study for Germany. ZEW-Centre for European Economic Research Discussion Paper, (18-033).

European Commission, 2015. Big Data Can Show Research Impact. Horizon: EU research and innovation magazine. < https://horizon-magazine.eu/article/big-data-can-help-show-impact-research-projects-eu-commissioner-moedas.html>

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

Katz, J.S. and Cothey, V., 2006. Web indicators for complex innovation systems. Research Evaluation, 15(2), pp.85-95.

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

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

LeCun, Y. and Bengio, Y., 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), p.1995.

Kim, Y., 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Oslo Manual., 2018. Proposed guidelines for collecting and interpreting technological innovation data. OCDE: Statistical Office of the European Communities.