First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3140

Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community

Rietsch, Constant[a]; Beaudry, Catherine[a] and Héroux-Vaillancourt, Mikaël[a]
[a]Department of Mathematics and Industrial Engineering, Ecole Polytechnique de Montréal, Canada,

Abstract

In this exploratory study, we explore a methodology using a web mining technique to source data in order to analyse innovation and commercialisation processes in Canadian nanotechnology firms. 79 websites have been extracted and analysed based on keywords related to 4 core concepts (R&D, intellectual property, collaboration and external financing) especially important for the commercialisation of nanotechnology. To validate our methodology, we compare our web mining results with those from a classic questionnaire-based survey. Our results show a correlation between the indicators from the two methods of r=0.306 (p-value=0.007) for R&D, of r=0.368 (p-value=0.002) for IP, of r=0.222 (p-value of 0.071) for Collaboration and of r=0.222 (p-value=0.067) for external financing. We conclude that some of the data extracted by our web mining technique can be used as proxy for specific variables obtained from more classical methods.

Keywords: *Web-mining, Innovation, Commercialisation, Nanotechnology.*

1. Introduction

Data is often hard to come by, and firms are increasingly solicited to answer surveys and participate in interviews. In this paper, we explore a methodology using a web mining technique to source data and analyse innovation and commercialisation processes in Canadian nanotechnology firms and help to overcome surveys issues.

Public websites are generally freely available and provide relevant information about a firm's products, services, business models, R&D activities and so on. All this information can be mined by researchers to study innovation and technology management. The question is whether this information is reliable and whether there is enough to give a good portrait of a firm characteristics - can the content of a commercial website be used to identify various innovation characteristics of a company? And if so, can we validate this methodology with concrete evidence?

Nanotechnology-related firms are especially interesting because of their broad set of applications and business sectors. As enabling technology vectors of the 21st century (Siegrist et al., 2007), the vast majority of nanotechnology-related companies have a website that is regularly updated. Regularly updated websites have the advantage of displaying more accurate data than what can be found in governmental databases (Gök, Waterworth, and Shapira 2014). In this study we analysed and compared the commercialisation of nanotechnology in Canada using two different techniques.

The remainder of the article is organised as follows: Section 2 presents the theoretical framework around web mining and our hypotheses about nanotechnology innovation and commercialisation; Section 3 describes the data and survey-based methodology; Section 4 presents and analyses the results; and finally Section 5 presents our conclusion.

2. Theory and hypotheses

The use of Internet data has the advantage of not being in direct contact with the subjects of the study and would ensure a distance between them and the study. Thus, the subject is not led to adapt his behaviour to the study, as can be the case with questionnaires and interviews. These types of unobtrusive measures are suitable for research inquiring for real actions but are restricted by the access of such a given population (Webb et al. 1966). Usually, this type of study is less expensive compared to intrusive studies such as questionnaires and interviews, which require researchers to perform extensive data collection (Lee 2000).

Nowadays, more innovation studies tend to rely on online questionnaires that companies must complete themselves inducing the multiple bias related with this technique. According to Sauermann (2013), numerous studies about innovation that had based their data

collection process on these Internet surveys typically received low response rates (between 10 and 25%) affecting the results of analysis by non-response bias. These online questionnaires are often complex and time-consuming for business managers, which explains why such a low response rate can be found.

The concept of the exploration of Internet data can be explained by the way in which we retrieve information about companies via their websites to convert them into analytical data. The vast majority of companies working in high technological fields such as the ones using nanotechnology keep their website updated in order to inform potential customers and investors about the current activities of the company. Of course, the information is made available online by the companies themselves, which indicates the possibility of a strong self-reporting bias. However, this source of information access would be suitable for the study of emerging technologies such as nanotechnology (Gök et al. 2014). Furthermore, Youtie et al. (2012) note that small businesses tend to have smaller websites which would facilitate the handling of data. However, it is clear that companies do not disclose all strategic and business data on their websites as it is already the case with other available data sources such as scientific publications or patents. A successful web mining analysis would have several advantages over questionnaires, scientific publications and patents. To start with, the population covered by a study using a search of the Web (web mining) is very wide (Herrouz, Khentout, and Djoudi, 2013) in an area where questionnaire studies find few returns, particularly in the field of new technologies. Contrary to government data, the frequency of updates is high, even daily, in most cases (Gök et al. 2014). Thus the information contained in websites is perfectly suited to many possible types of studies in the field of new technologies. The main disadvantage is the difficulty to organise and interpret data, with each site having different information and being organised differently.

In this study, we focus on parameters influencing innovation and commercialisation of the nanotechnology of Canadian firms. Based on Lee et al (2013), 4 important factors are considered to especially influence the commercialisation of nanotechnology: R&D, intellectual property, collaboration and external financing. Innovation and R&D efforts are likely to influence positively the firms' commercialisation and financial performance as mentioned in many studies (Geroski et al. 1993; Klette and Griliches, 2000). For nanotechnology firms, R&D efforts are likely to give them a technological superiority on the market. Intellectual property, especially patents, are the research outputs giving the company a competitive advantage over the competition by providing the exclusive product research for commercialisation. Technology patenting implies a return on investment by marketing the technology, reselling the patent or selling licenses. Moreover, patent statistics are also often use as a proxy for innovative activities (Pavitt 1985). Collaboration is essential for the development and the deployment of emerging technologies. McNeil et al. (2007) show that collaboration with universities or government institutes allows young companies to access especially expensive tools. Furthermore, Kim et al. (2008) stress the

impact of university research and scientists in the industry by providing specialized manpower, patents and innovation. Finally, most nanotechnology projects are still in their early stages, meaning they need private or public funding to attain the commercialisation phase. Most SMEs require public funding or venture capital investment to support nanotechnology commercialisation helping them to bridge the valley of death (Kalil, 2005; McNeil et al., 2007).

R&D, intellectual property, collaboration and external financing have all synonyms and other related terms that a company can use to refer to it. When we visit a company's website, we are directed to read what the company wants us to read. Companies use words that can give insight into what they actually do. We suggest that the more a company uses terms related to a certain factor, the more they are likely to perform activities related to that specific factor. Thus, from the 4 factors we mentioned earlier, we suggest the 4 following propositions:

Proposition 1: The more words related to R&D are used on a firm's website, the more a firm would be likely to perform R&D activities.

Proposition 2: The more words related to intellectual property are used on a firm's website, the more a firm would be likely to perform intellectual property related activities.

Proposition 3: The more words related to collaboration are used on a firm's website, the more a firm would be likely to perform collaborations.

Proposition 4: The more words related to external financing are used on a firm's website, the more a firm would be likely to perform external financing activities.

Each proposition will be tested with the help of the results of a classic questionnaire-based survey using the methodology explained in the following section.

3. Methodology

3.1. Data collection and sample methodology

We started by conducting a classic questionnaire-based survey of which the core is based on the Oslo Manual (OECD and Eurostat, 2005) and explored the following themes: innovation, commercialisation, collaboration and intellectual property. A sample of the questionnaire can be found in Annex I.

Firms that either use or develop nanotechnology are not labelled nor searchable in any obvious way. We used a list of 583 firms from AGY consulting, a Canadian firm specialized in emerging technologies such as nanotechnology, clean technology and biotechnology. We asked the companies whether they were performing nanotechnology

activities or using nano-enabled products or processes. When the companies were eligible to the study, we listed them with their associated NAICS code. We used a total of 23 NAICS codes representing 67% of all the cumulated frequencies with which we bought lists of over 3000 companies. We thus contacted 2971 high technological Canadian firms. 973 firms did not respond, 1439 were not eligible to the survey, 380 refuses to participate and a total of 222 were eligible. The first 13 fully-completed questionnaire served to test and validate the questionnaire in order to mitigate any self-reporting and fatigue bias. We did remove 6 questions in order reduce the time of completion and reduce potential fatigue bias. A total of 89 respondents finally accepted to participate to our study allowing us to reach a response rate of 40%. Since the population is unknown, we are in a presence of a non-probabilistic convenience sample for which it is possible the methodology induced a selection bias. Of course, we assume that the respondents were honest and answered the survey with goodwill.

Our sample represent a wide range of Canadian nanotechnology firms. Moreover, 74 % of the firms are nanotechnology intensive, which means that at least 80% of their revenues come from nanotechnology-related innovations. The different application domains in nanotechnology are wide with 54% for advanced materials, 21% for biotechnology and medicine, 24.4% for electronics, 23.30% for equipment and devices, 13.3% for photonics and 33.3% for other. More than 50% of respondents are small businesses and 83.5 % are SMEs with an average of $94 M revenues and $31M without the 3 biggest firms. Finally, 85% of the firms came from Quebec and Ontario and 12 % are from British Columbia and Alberta.

In order to test several types of bias such as self-reporting bias, non-respondent bias and non-selection bias, we gathered 79 eligible enterprises that did not participate to the study into a control sample. To do so, we needed an external source of data to validate our main sample. Industry Canada provides a database of companies in different sectors. The database is comprised of data provided by the companies themselves on a voluntary basis. While Industry Canada does not guarantee the accuracy or the reliability of the content, we assumed the companies that willingly updated information in an official public database will input accurate information and thus mitigate the self-reporting bias from this source. We used the data available from Industry Canada where we found the number of employees for 37 firms and revenues for 30 firms from our main sample and the number of employees for 29 firms and revenues for 26 firms from our control sample. We compared our main sample and our control sample with these two metrics with a Mann-Whitney U test and we did not find a significant difference between both samples for both metrics (p-value=0.115, p-value=0.166) which leads us to assume we are not likely to face a non-respondent bias.

We then compared these two metrics between the data obtained via our questionnaire-based survey and the data from Industry Canada in order to verify if any important self-reporting

bias can be found. For every firm for which we had both data from our Questionnaire and from Industry Canada, we tested each data pair with a Wilcoxon Signed Ranks Test. We did not find any significant differences between our questionnaire results and the data from Industry Canada (p-value=0.058, p-value=0.714), which leads us to assume the self-reporting bias issued from the questionnaire is not different from the one we can find in an official public database.

3.2. Web mining methodology

Next we selected these 89 enterprises, and used a web scraper, Nutch, to extract and store the text from their website. Due to technical limitations such as the structure of the websites, only 79 of these firms (88%) provided enough information to be included in our study. We then used a content mining technique to perform a word frequency analysis with the text present on the websites. More specifically, in the 79 websites, we looked for innovation and commercialisation core factors : R&D, intellectual property, collaboration and external financing. For each factor, we listed all the relevant keywords that appear in company web pages. Factors, keywords and the web mining construct are described in Annex II. R&D and collaboration keywords were selected from the literature while intellectual property and external financing are issued from our own research. The Government of Canada offers many public programs and funding opportunities to companies for the development of nanotechnology projects. The website of Industry Canada identifies funds and programs offered to Canadian nanotechnology firms that we have used for our research.

Clustering using keyword frequency analysis with a text mining software enabled us to get the occurrences of each keyword for each factor. We transformed these clusters of occurrences into 4 continuous variables. Because the 79 companies are different in structure and size and therefore, present different amounts of information in their websites, we standardized each variable by dividing all occurrences by the total number of words appearing on their website and multiplied the resulting value by 1000. For each continuous variable, we obtained the Kurtosis and Skewness measures in order to determine whether our variables were following a normal distribution. All 4 variables did not follow a normal distribution so we transformed them by applying a natural logarithm (LN) or an inverse function (INV). In the case of External financing, we did not reach normality and thus, we treated this variable with a non-parametric test.

Since we selected only the companies that answered the survey, there is a possible selection bias. We ran our web mining technique on our control sample and generated the same variables. We then used a Student's t-test to test the difference of means for the following variables LN_WEB_MINING_RD (p-value=0.13), INV_WEB_MINING_IP (p-value=0.083) and LN_WEB_MINING_COLLAB (p-value=0.144) and conclude that the

difference is not significant between the two sample for these three variables. We tested the means of the variable WEB_MINING_EXTERN_FINAN (p-value=0.008) with a Mann-Whitney U test and found it was significant, so we cannot conclude for that variable that the means of the two sample are the same. Therefore, a selection bias is present for WEB_MINING_EXTERN_FINAN and will be included in our limits of research.

3.3. Questionnaire-based survey data methodology

In order to validate our 4 continuous variables from our web mining results, we identified all the relevant questions from the questionnaire-based survey and transformed them into different types of variables. The questions used can be found in Annex 1. We transformed every continuous variables from the survey that did not follow a normal distribution by applying a natural logarithm (LN) or an inverse function (INV). Since several 7-point Likert scale questions described the concept of R&D, we used Principal Component Analysis (PCA) with a Varimax rotation to reduce the number of variables and combine them into relevant dimensions corresponding to specific factors of the concept examined. Two factors were created but both the K-M-O and Cronbach alpha did not reach an acceptable level which would satisfy the validity and the reliability of the construct. In addition, these combined variables do not correlate with each other which hints towards using a formative construct. We thus proceeded to treating each item individually.

At the end, we generated a total of 9 variables corresponding to R&D, 1 variable related to collaboration, 2 variables corresponding to external financing and finally, 2 variables measuring intellectual property. The details of the Questionnaire-based survey construct can be see in Annex III.

3.4. Web mining validation with questionnaire-based survey methodology

Each pair of variables related to the same concept from the two methods (Web mining and survey) was examined via a Pearson correlation analysis when the subjects were following a normal distribution or a Spearman correlation when they were not following a normal distribution, to assess whether the variables stemming from the Web mining analysis can be used as a proxy for similar concepts measured by a survey. The details concerning our construct comparing a Web Mining technique and a Questionnaire-based survey can be see in Annex IV.

4. Results

The results of this paper aim to validate the utilisation of a web-mining-based methodology using firms' websites as a data source to analyse the extent of commercialisation and innovation, which can be used to better understand innovation practices. Comparing the variables constructed from the web mining and from the survey, we find a correlation of

0.306 (p-value of 0.007) between R&D measures and whether a firm is likely to provide R&D services to third parties. Additionally, we find a correlation of 0.306 (p-value of 0.010) when we associate the R&D concepts on websites with whether a firm has a high percentage of employees allocated to R&D tasks. Moreover, we find a correlation of 0.284 (p-value of 0.013) when we associate the R&D concepts on websites with whether a firm is likely to contract R&D service from external providers. Finally, we find a non-significant correlation of 0.197 (p-value of 0.100) when we associate the R&D concepts on websites with whether a firm has a long R&D process or not. It is important to note that the variable LN_NUMBER_RD which is related to the number of R&D projects did not correlate at all with our web mining variable with r =0.002 (p-value=0.985).

Terms related to intellectual property strongly correlate with the variables from the survey with a correlation of 0.368 (p-value of 0.002) regarding the use of intellectual property mechanisms and with a correlation of 0.351 (p-value of 0.033) regarding the activities related to patenting. Web mining methods therefore appear to be able to capture the importance of the use of IP mechanisms.

Collaboration terms from the Web sites are partially correlated with r=0.222 (p-value of 0.071) with the firms that confirmed collaborating from our questionnaire but the result is not significant at 5%.

External financing terms (from the web-based analysis) are also partially correlated with the extent of the use of external funds for commercialisation purposes (r=0.222 – p-value of 0.067) but the result is not significant at 5% regarding their importance for funding R&D activities.

To conclude, our latest results confirm the data extracted by our web mining technique can be used as a proxy at least for some of the variables coming from classical methods. If the collaboration and financing concept did not have a significant correlation, intellectual property and most of R&D web mining variables seem to be, according to our findings, good proxies for innovation studies.

5. Discussion and conclusion

Websites are a gold mine of informations. Researchers in innovation and technology management are now investigating if they can datamine enterprises' websites in order to get valuable data to their research. Nowadays, researchers rely on questionnaire-based survey to get most of the data. These questionnaires are costly, time consuming and a source of multiple bias. We thus explore a technique using data mining to determine if whether or not we can use data from websites as proxy for certain information that would have required a questionnaire-based survey to be obtained. We tested 4 factors that are determinant for the success of nanotechnology commercialisation: R&D, intellectual property, collaboration

and external financing. While results seems conclusive for intellectual property factors and some indicators of R&D, results did not show significant correlation with neither collaboration and external financing factors. Therefore, our proposition 3 and 4 are not yet validated.

For the specific case of R&D, we can observe that the web mining indicator seems to reflect the promotion needs in terms of R&D. Our web mining R&D indicator did correlate the most when firms are more likely to provide R&D services to third parties. This might be explained by the fact that the company use its website to promote their offer of R&D service. Also, our web mining R&D indicator did correlate really high with firm has a high percentage of employees allocated to R&D tasks. This can be explained by the willingness of a firm to attract new talents in R&D through their websites. Finally, our web mining R&D indicator did correlate significantly when firms are more likely to contract R&D service from external providers. However, one of the most important R&D indicator, the number of R&D projects, did not correlate at all with our web mining variable which may seems counter intuitive. Thus, we can hardly use our web indicator as a proxy since 3 independent indicators correlates with it and strong indicator of R&D activities are ignore. Therefore, our proposition 1 is partially true. A better definition of R&D activities would be required in order to use a R&D web mining proxy.

Our intellectual property web mining correlates with both the use of intellectual property mechanisms and the activities related to patenting. In that sense, our second proposition seems to be true i.e. intellectual property activities seems it can be explained by an IP web mining proxy.

More data would allow our research to be more robust, especially when it comes to verifying the concept of collaboration and external financing, normally addressed with classical methods, can be appropriately measured on web sites. For instance, we were not able to crawl data from all the companies from our survey due to technical limitations and only 79 out of 89 companies were used in this paper. Another limitation of our methodology is that we did not take into account the context of our keywords, possibly leading to multiple false positives. For instance, the mention of the word ‘collaboration’ on a website does not necessarily means that the company does collaboration with second parties at all. Qualitative data analysis of the websites’ content could be used to reduce the risk of false positives and to gather more accurate data. Moreover, our data are limited to textual content, while website also display, images, sounds and videos which are difficult to take into account in our study. Of course, websites, questionnaire-based survey and the official public database we used are all subject to self-reporting bias and it is part of our limitations.

Websites can be updated from time to time and the results can change accordingly depending on what companies want to display publicly. Thus, it is important to note that a

punctual web mine crawl might not be sufficient to capture all relevant information and results are subject to change with updated websites. Thus, longitudinal study would be required to better assess the validity of our methodology over time.

In the very near future, Partial Least Square (PLS) regression will be tested to determine if it is possible to create reliable and valid reflective indexes from the factors found by the PCA. In addition, we are currently investigating the use of a Multitrait-multimethod matrix (MTMM) to verify the validity and reliability of our constructs and to determine whether our methodology can be used as a valid approach to provide data for future innovation and technology management studies. Future studies will allow to better understand whether these web mining indicators capture all the information required to understand the proposed factors and can be used as a proxy for questionnaire-based survey questions or if these variables propose additional information that was not captured before by traditional means.

References

Geroski, Paul, Machin, Steve & Van Reenen, John. (1993). "The profitability of innovating firms." *The RAND Journal of Economics*, 198-211.

Gök, Abdullah, Alec, Waterworth, & Philip Shapira. (2014). "Use of Web Mining in Studying Innovation." Scientometrics, September, 1–19. doi:10.1007/s11192-014-1434-0.

Herrouz, Abdelhakim, Chabane Khentout, & Mahieddine Djoudi. (2013). "Overview of Web Content Mining Tools." *arXiv Preprint arXiv:1307.1024*. http://arxiv.org/abs/1307.1024.

Kalil, Thomas A. (2005). "Nanotechnology and the valley of death." 265.

Kim, Jinyoung, Sangjoon John Lee, & Gerald Marschke. (2008). "Impact of University Scientists on Innovations in Nanotechnology."

Klette, Tor Jakob, Jarle Møen, & Zvi Griliches (2000). "Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies." *Research Policy* 29, no. 4, 471-495.

Lee, C. J., Lee, S., Jhon, M. S., & Shin, J. (2013). Factors influencing nanotechnology commercialization: an empirical analysis of nanotechnology firms in South Korea. Journal of nanoparticle research, 15(2), 1-17.

OECD and Eurostat. (2005). Oslo Manual. The Measurement of Scientific and Technological Activities. OECD Publishing. http://www.oecd-ilibrary.org/science-and-technology/oslo-manual_9789264013100-en.

Pavitt, K. (1985). "Patent Statistics as Indicators of Innovative Activities: Possibilities and Problems." *Scientometrics* 7 (1-2): 77–99. doi:10.1007/BF02020142.

Ramdani, Anas, Vue de, Du diplôme de maîtrise l'obtention, and others. (2014). "revue systématique de la littérature sur les mesures de la collaboration inter-organisationnelle dans un contexte d'innovation." http://publications.polymtl.ca/1624/1/2014_AnasRamdani.pdf.

Sauermann, Henry, & Michael Roach. (2013). "Increasing Web Survey Response Rates in Innovation Research: An Experimental Study of Static and Dynamic Contact Design Features." Research Policy 42 (1): 273–86. doi:10.1016/j.respol.2012.05.003.

Siegrist, Michael, Carmen Keller, Hans Kastenholz, Silvia Frey, & Arnim Wiek. (2007). "Laypeople's and experts' perception of nanotechnology hazards."*Risk Analysis* 27, no. 1, 59-69.

Webb, E. J., Campbell, D. T., & Schwartz, R. D. (1966). Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally.

Youtie, Jan, Diana Hicks, Philip Shapira, & Travis Horsley. (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): 981–95. doi:10.1080/09537325.2012.724163.

Annex I - Questions from the questionnaire-based survey

R&D

1- How many nanotechnology-related and/or advanced material products in development do you actually have in each of the following phases?

1- Applied Research, 2- Product Scoping and Business Case Building, 3- Development, Testing and Validation, 4- Commercialisation

2- How important to your plant's innovation activities are each of the following sources of knowledge and innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Internal R&D in your firm
- Commercial laboratories / R&D firms / Technical Consultants

3- Please indicate the level of importance of each of the following innovation activities to your plant during the period 2010 to 2014 (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Contracting of external R&D service providers
- Providing R&D services to third parties

4- How long did it take to develop your most significant and recent (MSR) nanotechnology-related product innovation?

5- How important were each of the following organisations as collaborators in the development and commercialization of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Private research laboratories / Research and Development firms

6- How important were the following reasons in deciding to collaborate for the development and the commercialisation of your MSR product innovation? (1-Not important, 2-Very Low, 3-Low, 5-High, 6-Very high, 7-Essential)

- Accessing research and development

7- What proportion of Canadian employees from your firm are assigned primarily in R&D (%)?

Collaboration

1- Did your firm participate in alliances or collaborative agreements with other organisations to develop or commercialise your MSR product innovation? Y/N

2- How important were each of the following organisations as collaborators in the development and commercialisation of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Universities or higher education institutions, College centres for technology transfer (CCTT) and CEGEPs, university technology transfer offices

External financing

1- Please indicate the proportion (%) of the total amount of financing provided by each of the following sources for the development and commercialisation of your MSR product innovation.

Y: 1- Internal funds of your firm or establishment, 2- Government subsidies / tax credits / academics grants, 3- Debt capital (such as bank loans), 4- Venture capital (public/private), 5- Collaboration agreements, 6- Programs from organisations such as nanoQuebec (now PRIMA-Quebec), nanoOntario, nanoAlberta, etc.., 7-Other

X: 1- Development of innovation, 2- Commercialisation of innovation

Intellectual property

1- Which of the following mechanisms are used by your firm to protect the intellectual property rights (IPR) for your MSR product innovation?

- Patents
- Trademarks
- Confidentiality agreements
- Trade secrets
- First mover advantage
- Other

2- How many patents does your firm own? Please note that the same patent filed in different countries is considered as only one patent.

Y: 1- Patent applications, 2- Existing patents, 3- Patents assigned / (sold) to others

X: 1- All patents, 2- Nanotechnology-related and advanced materials patents

Annex II - Web Mining construct

Topic	Factors	Keywords	Indicators	Variables
Innovation and commercialisation of nanotechnology (Lee et al, 2013)	R&D	research and development, r&d, laboratories, researcher, scientist, product development, technology development, development phase, technical development, development program, development process, development project, development cent, development facility, technological development, development effort, development cycle, development research, research & development, development activity, fundamental research, basic research (Gök et al. 2014)	Number of keywords frequencies per webpage	LN_WEB_MINING_RD (Continuous, normal)
	Intellectual property	Patent, intellectual property, trade secret, industrial design	Number of keywords frequencies per webpage	INV_WEB_MINING_IP (Continuous, normal)
	Collaboration	affiliation, collaboration, cooperation, partners, partnership (Ramdani et al. 2014)	Number of keywords frequencies per webpage	LN_WEB_MINING_COLLAB (Continuous, normal)
	External financing	atlantic canada opportunities agency, business development bank of canada, sustainable development technology, venture capital, atlantic innovation fund, nrc-irap, fednor, Industrial research assistance program, grants, private investment	Number of keywords frequencies per webpage	WEB_MINING_EXTERN_FINAN (Continuous, not normal)

Annex III - Questionnaire-based survey for web mining validation construct

Concepts	Indicators	Variables
R&D	• Number of R&D projects in nanotechnology • Level of importance of internal R&D as a source of knowledge • Level of importance of Commercial laboratories / R&D firms / Technical Consultants as a source of knowledge • Level of importance of contracting of external R&D service providers • Level of importance of providing R&D services to third parties • Time of R&D • Level of importance of Private research laboratories / Research and Development firms as collaborators for the development and the commercialisation • Level of importance of accessing research and development from collaborators for the development and the commercialisation • Proportion of Canadian employees assigned primarily in R&D (%)	• LN_NUMBER_RD (Continuous, normal) • D_INTENSITY_INTERN_INFO_RD (Dummy) • INTENSITY_EXTERN_INFO_RD (Continuous, normal) • INTENSITY_CONTRACTING_RD (Continuous, normal) • INTENSITY_PROVIDING_RD (Continuous, normal) • LN_TIME_RD (Continuous, normal) • D_INTENSITE_COLLAB_RD (Dummy) • D_INTENSITE_COLLAB_REASON_RD (Dummy) • PROP_RD (Continuous, normal)
Intellectual property	• Number of IP mechanisms used • Number of patents	• SUM_IP (Continuous, normal) • LN_NUMB_PATENT (Continuous, normal)
Collaboration	• Use of collaboration for the latest innovation	• D_COLLAB (Dummy)
External financing	• Proportion of external financing for R&D (%) • Proportion of external financing for commercialisation (%)	• RD_EXTERN_FINAN (Continuous, normal) • COMM_EXTERN_FINAN (Continuous, normal) • TOTAL_EXTERN_FINAN (Continuous, normal)

Annex IV - Questionnaire-based survey for web mining validation construct

Concepts	Web Mining Variables	Questionnaire Variables	Correlation type
R&D	LN_WEB_MINING_RD (Continuous, normal)	• LN_NUMBER_RD (Continue, normal) • D_INTENSITY_INTERN_INFO_RD (Dummy) • INTENSITY_EXTERN_INFO_RD (Continue, normal) • INTENSITY_CONTRACTING_RD (Continue, normal) • INTENSITY_PROVIDING_RD (Continue, normal) • LN_TIME_RD (Continue, normal) • PROP_RD (Continue, normal) • D_INTENSITE_COLLAB_RD (Dummy) • D_INTENSITE_COLLAB_REASON_RD (Dummy)	Pearson
Intellectual property	INV_WEB_MINING_IP (Continuous, normal)	• SUM_IP (Continuous, normal) • LN_NUMB_PATENT (Continuous, normal)	Pearson
Collaboration	LN_WEB_MINING_COLLAB (Continuous, normal)	• D_COLLAB (Dummy)	Pearson
External financing	WEB_MINING_EXTERN_FINAN (Continuous, not normal)	• RD_EXTERN_FINAN (Continuous, normal) • COMM_EXTERN_FINAN (Continuous, normal) • TOTAL_EXTERN_FINAN (Continuous, normal)	Spearman