RESEARCH ARTICLE

Using web content analysis to create innovation indicators—What do we really measure?

Mikaël Héroux-Vaillancourt, Catherine Beaudry, and Constant Rietsch

Canada Research Chair on the Creation, Development and Commercialization of Innovation, Department of Mathematics and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Downtown office Montreal, Quebec, H3C 3A7, Canada

Keywords: construct validity, innovation measurement, multitraits multimethods, web content analysis, web-mining, word frequency analysis

ABSTRACT

This study explores the use of web content analysis to build innovation indicators from the complete texts of 79 corporate websites of Canadian nanotechnology and advanced materials firms. Indicators of four core concepts (R&D, IP protection, collaboration, and external financing) of the innovation process were built using keywords frequency analysis. These web-based indicators were validated using several indicators built from a classic questionnaire-based survey with the following methods: correlation analysis, multitraits multimethods (MTMM) matrices, and confirmatory factor analysis (CFA). The results suggest that formative indices built with the questionnaire and web-based indicators measure the same concept, which is not the case when considering the items from the questionnaire separately. Web-based indicators can act either as complements to direct measures or as substitutes for broader measures, notably the importance of R&D and the importance of IP protection, which are normally measured using conventional methods, such as government administrative data or questionnaire-based surveys.

1. INTRODUCTION

The majority of researchers in innovation and technology management today still rely on public databases and questionnaire-based surveys to obtain most of their data to perform quantitative studies on industrial strategies and innovation activities. However, public databases are often incomplete or too general in nature. Although questionnaires remain precise instruments, the process of designing, testing, and administering questionnaire-based surveys can be especially time-consuming and costly for researchers. Furthermore, oversolicitation of respondents and of their time militates against questionnaire-based surveys, which suffer from increasingly lower response rates (less than 10%) and thus threaten the validity of studies performed using such methods, for instance because of the potential for significant nonresponse biases.

To complement questionnaire-based data, social scientists have often used secondary data. With the development of "Big Data" analytical tools, websites are increasingly recognized as additional information gold mines. Researchers in innovation and technology management are now investigating whether they can use the information that organizations provide on their websites to acquire valuable additional data for their research. Technology companies generally maintain websites that allow the media, as well as potential investors, customers, suppliers, and collaborators, to learn about the nature of their activities. The information provided on company websites is as rich as it is diversified, including products, services, business models, R&D activities, and more. As these corporate websites are freely available to anyone with internet access,

an open access journal

Check for updates

Citation: Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators—What do we really measure? *Quantitative Science Studies*, *1*(4), 1601–1637. https://doi.org/10.1162/qss_a_00086

DOI: https://doi.org/10.1162/qss_a_00086

Received: 26 March 2019
Accepted: 28 May 2020

Corresponding Author: Mikaël Héroux-Vaillancourt
mikael.heroux-vaillancourt@polymtl.ca

Handling Editor: Vincent Larivière

The MIT Press

researchers need to evaluate their value as data sources. Before fully validating their usage, a few questions need to be answered: Is it possible to extract the information contained on these websites and convert it into useful data for research purposes? Is the available information reliable and sufficient to provide an accurate picture of the firm's specific characteristics? In other words, can the contents of a corporate website be used to identify different business innovation attributes?

This research aims to study high technologies that should have emerged by now, nanotechnologies and advanced materials, but that are still difficult to assess because of their pervasiveness throughout the industry. Nanotechnologies and advanced materials, as so-called enabling technologies, have innumerable applications and are present in all major industrial sectors, such as food, agriculture, electronics, renewable energy, the environment, biomedical, and healthcare. The ability to develop and use nanotechnologies and advanced materials is expected to give impetus to innovation performance and economic growth (Hwang, 2010). The versatility of these advanced technologies' applications and the sectors affected by them make them particularly interesting for studies in innovation management because they allow an intersectoral bird's-eye view of the industry. However, owing to the pervasiveness of nanotechnologies and advanced materials throughout the economy, national statistics and other data repositories often fail to accurately measure their impact. A potential solution to alleviate this perceived lack of accurate data lies in the information contained on company websites.

The clear majority of companies working in high-technology fields, such as those producing or using nanotechnology and advanced materials, maintain updated websites. Although the online information is made available by the companies themselves, which suggests the possibility of a strong self-reporting bias, this source of information is deemed suitable for the study of the emerging technologies (Gök, Waterworth, & Shapira, 2015). For instance, Youtie, Hicks, et al. (2012) noted that small businesses tend to have less content-heavy websites, which facilitates the handling of data. A successful web mining analysis has several advantages over questionnaires, scientific publications, and patents. First, the population covered by searching the web (web crawling) is very large (Herrouz, Khentout, & Djoudi, 2013) compared to questionnaire-based studies, which generate few returns (low response rate). Contrary to government data, the frequency of updates is high, even daily, in most cases (Gök et al., 2015). One significant drawback, however, is that companies do not disclose all of their strategic and business data on their websites. The main disadvantage of such web-based data stems from problems inherent in organizing and interpreting highly unstructured information, as each site organizes various types of information differently. In addition, we do not yet know exactly what the significance is of what we are measuring, hence the necessity to concurrently use both traditional and webometrics methods and to compare them with one another.

In this study, we analyzed and compared four sets of measures of innovation and commercialization respecting nanotechnology and advanced materials firms in Canada stemming from two different data gathering techniques: word frequency analysis from web content and questionnaire-based surveys. Comparisons between the results from both methods were obtained via correlations. To ensure a convergent and discriminant validation of our results, a multitraits multimethods (MTMM) technique was then performed on the most significant measures. Formative indices were built to determine the best representation of the web-based indicators. One final MTMM matrix was used, along with an MTMM confirmatory factor analysis (CFA), as a post hoc analysis to ensure the robustness of our results.

The remainder of the article is organized as follows: Section 2 presents the theoretical development of the innovation and commercialization of high technologies, such as nanotechnology and advanced materials, along with web content analysis; section 3 describes the methodologies

used to collect data and the MTMM method; section 4 analyzes the results surrounding propositions 1 and 2, and performs a post hoc analysis; section 5 discusses the results and conclusions; and lastly, section 6 addresses the research limitations and future research.

2. THEORETICAL DEVELOPMENT

2.1. Using the Web as a Data Source for Social Sciences

A few years after the invention of the world wide web, its use as a means to get valuable data for social science studies was introduced (Almind & Ingwersen, 1997). With more and more online content from all aspects of society being produced, the web becomes a digital representation of the social construct that embodies our civilization. With such a rich repertoire of information, social scientists started to explore ways to leverage the web as a data source for research (Björneborn & Ingwersen, 2004). As a result, measuring different elements of the web, such as websites, webpages, parts of webpages, words in webpages, hyperlinks, and web search engine results, has been undertaken by a number of scholars, paving the way to a new field of research: webometrics, as an evolution from the classic bibliometric studies (Thelwall, 2009).

Link analysis, blog searching, and web impact assessment were the main focus of late 2000s webometrics research. The first, link analysis, consists in analyzing the network of hyperlinks contained in web pages in the hope that they may reflect an organization's connections (Hyun Kim, 2012; Katz & Cothey, 2006; Vaughan, 2004). This method has proven useful to understand the structure of the collaborative network and the importance of a given actor within its network (Minguillo & Thelwall, 2012; Stuart & Thelwall, 2006). The interest in the second method, exploring the content of blogs, resides in their vast heterogeneous public-generated information at a given time. Their understanding provides information on the trends in public opinion on a given set of topics. Nowadays, blog searching is more prevalent in social media platforms such as Twitter (Thelwall, Buckley, & Paltoglou, 2011), and the importance of topics can be effectively monitored through Google Trends (Choi & Varian, 2012). The third, web impact assessment, offers a method for measuring the direct online impact of specific topics or documents by counting their presence on the web. These could be a good proxy for offline impact due to the dominant penetration of the internet as a means of information and communication (Thelwall, 2009). Web impact assessment could help identify patterns of diffusion and relative impact of authors, terms, scientific theories, political candidates, books, journals, and so on. In the case of innovation studies, the web could be viewed as a way to obtain indirect indicators, such as the number of attendees at a given event, the media coverage for a given product launch, or the number of requests for an organization's literature or the number of times an organization's concepts or publications are mentioned (Thelwall, 2009).

To perform web analysis, three main areas of web mining are generally used: web structure mining for link analysis; both web usage mining and web content mining for web impact assessment; and blog searching (Miner, Elder, et al., 2012). Of particular interest for this paper is web content mining, which is a technique that turns unstructured information in the form of text contained in web pages into structured data useful for research purposes.

Content analysis is an objective and quantitative method used to find relationships between textual information and the context from which the information is sourced (Krippendorff, 1980). The benefits and challenges of using the web as a data source for content analysis were already discussed at the beginning of the millennium (Weare & Lin, 2000). The main opportunity from using the web as a data source stems from the explosion of possible sources of multimedia information (text, video, audio, and image) that can be used for content analysis, which was anticipated to reduce the cost of data acquisition significantly. However, the information contained on

the web is highly unstructured, completely decentralized, and not standardized in any way, which makes the quality and validity of web content analysis particularly challenging to verify (Weare & Lin, 2000) and somewhat costly to process.

Despite the limitations, some web content analyses were performed in innovation studies. For example, Herrouz et al. (2013) used a web scraping method to obtain data from small and medium-sized high-technology graphene firm websites in the United States, United Kingdom, and China. More recently, Gök et al. (2015) proposed a web content analysis based on a keywords frequency analysis to assess the R&D activities of 296 UK-based green goods small and midsize enterprises (SMEs). They compared the results of their R&D web indicator with the results obtained through a questionnaire-based survey. The results of this study showed that the web-based indicator did not correlate significantly with the nonweb-based indicator, which suggests that these two indicators did not reflect the same concept. It is therefore possible that web-mining indicators provide new information that was not captured through classical methodologies and thus are shown to be complementary.

2.2. Construct Definition

Following in the footsteps of Gök et al. (2015), in this study we focus on parameters influencing innovation and commercialization in Canadian high-technology firms. We consider the four most important factors known to influence innovation and commercialization in high-tech firms, in our case, nanotechnologies and advanced materials: research & development (R&D) intensity, intellectual property (IP) protection, collaboration, and external financing (Lee, Lee, et al., 2013).

R&D intensity refers to the effort a firm makes to generate inventions (Griliches, 1990, 1994, 1998; Hausman, Hall, & Griliches, 1984; Hitt, Hoskisson, & Kim, 1997) as well as investment in its own absorptive capacity (Cohen & Levinthal, 1990). Greater R&D efforts are likely to yield better returns to innovation. Indeed, the positive relationship between R&D efforts and innovation performance has been demonstrated in several quantitative studies (Baysinger & Hoskisson, 1989; Deeds, 2001; Greve, 2003; Griliches, 1998; Hagedoorn & Cloodt, 2003; Hall, 1990; Parthasarthy & Hammond, 2002). Innovation and R&D efforts have been shown to positively influence firms' commercialization and financial performance (Geroski, Machin, & Van Reenen, 1993; Klette, Møen, & Griliches, 2000). R&D efforts should therefore give nanotechnology and advanced materials firms a technological superiority as compared to the market. R&D intensity is a measure of intention, and, as such, not a direct measurement of the performance of the R&D process as concerns the production of knowledge, inventions, and innovations (Adams, Bessant, & Phelps, 2006; Cebon, Newton, & Noble, 1999; Flor & Oltra, 2004; Kleinknecht, Van Montfort, & Brouwer, 2002). In addition, not all innovations are systematically derived from this internal R&D process (Michie, 1998), as highlighted by the rising popularity of open innovation practices (Chesbrough, 2003). Moreover, measures based solely on R&D are not the best suited for SMEs because their efforts are often neither formal (Kleinknecht et al., 2002) nor constant (Michie, 1998). For these reasons, researchers use R&D intensity less as a proxy for innovation, as suggested by (Becheikh, Landry, & Amara, 2006), and more as a means to achieve higher innovation performance.

IP protection confers a competitive advantage to companies by offering exclusivity for the commercialization of technologies, branding, copyright, industrial design, and so on. (Teece, 1986). Patents enhance the return on investment of the amounts dedicated to technology development through temporary monopoly power, license granting, or the sale of patents (Arora, 1995; Chesbrough, 2003; Feldman & Florida, 1994; Fosfuri, 2006; Mazzoleni & Nelson, 1998; Merges, 1999; Rivette & Kline, 2000). Laursen and Salter (2006) observed a positive relationship between

appropriability and innovation performance, but one that eventually leads to decreasing returns. A patent, as one of the outputs of the research process, is often considered as a measure of inventiveness (Coombs, Narandren, & Richards, 1996; Flor & Oltra, 2004; OECD & Statistical Office of the European Communities, 2005). Although patent statistics are frequently used as proxies for innovative activities (Pavitt, 1985) or innovation (Becheikh et al., 2006), they have also been the subject of considerable criticism (Archibugi, 1992; Cohen & Levin, 1989; Dosi, 1988; Griliches, 1998). The use of patents varies considerably from one sector to another (Archibugi & Sirilli, 2001; Armellini, Beaudry, & Kaminski, 2017; Armellini, Kaminski, & Beaudry, 2014; Michie, 1998) and according to firm size, because SMEs, which have fewer financial resources to protect their IP, are systematically disadvantaged.

In contrast to the first two types of indicators, collaboration has a much broader impact on the innovation process: from idea generation and research to commercialization. Collaboration thus has a positive impact on several innovation performance indicators, such as the number of patents, sales growth, and the return on innovation sales (Arvanitis, 2012; Belderbos, Carree, & Lokshin, 2004; Carboni, 2013), and has been shown to improve competitiveness (Hagedoorn, Link, & Vonortas, 2000; Roja & Nastase, 2013). A constant collaborative effort is essential for the development and deployment of emerging technologies, which are becoming increasingly complex, as well as for reducing R&D costs, sharing risk, and increasing performance (Johnson & Filippini, 2009; Parker, 2000). For example, the aerospace industry requires all actors to adopt a policy of interorganizational cooperation to take advantage of all the necessary knowledge and know-how available (Jordan & Lowe, 2004). In this field, collaboration is so important that it spans all types of partners: clients (Armellini et al., 2017), including suppliers (Bozdogan, Deyst, et al., 1998); competitors (Esposito, 2004; Frear & Metcalf, 1995); and universities and institutes (Armellini et al., 2014, 2017). McNeil, Lowe, et al. (2007) have shown that collaboration with universities or government institutes enables young high-tech firms to access particularly expensive tools. In addition, Kim, Lee, and Marschke (2014) highlighted the impact of university research on scientists associated with the training and development of skilled labor, patents, and innovation in industry.

Finally, as most nanotechnology and some advanced material projects are still in the early development stages, they rely heavily on private and/or public funding to reach the commercialization and marketing phases (Kalil, 2005; McNeil et al., 2007). In surveys on barriers to innovation, firms very often refer to the lack of external financing as a major obstacle to their innovation activities (Harhoff & Körting, 1998). Venture capital investment in young U.S. high-tech firms had a positive and significant effect on their R&D productivity in the 1990s (Brown, Fazzari, & Petersen, 2009). For more than a decade, substantial public investments have targeted nanotechnologies worldwide (Crawley, 2007); more recently, in 2018, C$1.2 billion were invested in the United States (National Nanotechnology Coordination Office, 2017). As the development of these emerging technologies requires public support, the Innovation, Science and Economic Development Canada (ISED, formerly known as *Industry Canada*) website proposes 13 programs dedicated to funding the development of nanotechnologies and advanced materials. These public funds often have a lever effect and firms that receive such funding in grant form are more likely to also successfully raise private funding and receive better external financing offers (Meuleman & De Maeseneire, 2012).

These four pillars of the innovation process (R&D, IP, collaboration, and external financing) have been the focus of much literature and have been extensively measured by both researchers and national statistics offices. But do they transpire in corporate websites? A number of scholars have attempted to analyze web content from the information contained on companies' websites to build indicators and innovation metrics (Gök et al., 2015; Herrouz et al.,

2013). Yet understanding the "real" meaning of measures created from data gathered with web mining techniques is not a trivial task. Gök et al. (2015) showed that their web-based indicator based on R&D-related keywords did not correlate significantly with nonweb-based indicators, which suggested that these two indicators did not reflect the same concept. Therefore, the web mining indicator could represent new information that was not captured through classical methodologies and could possibly be complementary. In this paper, two propositions will be tested to assess whether these web-based indicators can be used as substitutes for indicators obtained by classical questionnaire-based methods.

Within corporate websites, innovation words and factors are referred to using several synonyms and other related terms. Companies may use words on their website to provide insight into what they actually do. The more a company uses terms related to a specific factor, the stronger the signal is deemed for that factor, and therefore, the more likely the firm is liable to perform activities related to that specific factor. Thus, for each factor mentioned above, we suggest the following proposition:

> *Proposition 1: The more words related to a factor, in our case* R&D intensity, intellectual property, collaboration, and external financing, *that are used on a firm's website, the more a firm is likely to perform activities related to that factor.*

Connecting keywords found on websites with tangible actions taken by those companies is a bold proposition that may not be valid empirically. The content of corporate websites is written with the aim of signaling the company's "best" characteristics to all its internal and external stakeholders. This represents a rather modest but reasonable claim against the validity of the information to be found, which may suffer from a possible self-reporting bias. However, it is reasonable to assume that the information provided on companies' websites will be as true to the company's intentions as to the various concepts it wishes to share with the world. The information that stands out on a website can give a broad sense of the importance of a specific factor for the firm.

> *Proposition 2: The more words related to a factor, in our case* R&D intensity, intellectual property, collaboration, and external financing, *that are used on a firm's website, the more important a firm considers that factor.*

The two propositions will be tested and compared to the results obtained with a classic questionnaire-based survey using the methodology explained in section 3.

3. DATA AND METHODOLOGY

3.1. Questionnaire-Based Data Collection

For a more general study of innovation in nanotechnologies and advanced materials, we first conducted a classic questionnaire-based survey, the core of which is based on the Oslo Manual (OECD & Statistical Office of the European Communities, 2005), to explore the following themes: innovation, commercialization, collaboration, and IP protection, among other topics. The questionnaire includes the importance of the sources of knowledge and the importance of innovation activities measured by a Likert scale from *not important* (1) to *essential* (7). Other questions cover the importance of commercialization actions, the proportion of financing for R&D and commercialization, the importance of market impacts, the number of exports, the importance of obstacles to commercialization, whether or not the firm collaborated to develop or commercialize the latest most significant product innovation, the importance of collaborators,

the importance of the reasons for collaboration, IP mechanisms, patent management, and general questions such as the number of employees, revenue, and the firm's business sector. A sample of the questionnaire is provided in Appendix 1.

Due to the ubiquitous nature of nanotechnology and advanced materials, combined with the fact that they are adopted in a large number of industrial sectors, identifying companies that use or develop these technologies is not straightforward. Firms that either use or develop nanotechnology and advanced materials are not labeled as such, nor are they searchable in any obvious way. As a consequence, we used all possible means to build an exhaustive list of high-tech companies with a higher probability to be involved in nanotechnology or advanced materials with the help of several sources, including ISED, Nano-Québec (now Prima-Québec), Nano Ontario, Nanowerk, futuremarketsinc.com, and AGY Consulting, a Canadian firm specialized in emerging technologies such as nanotechnology, clean technology, and biotechnology. Combining all these data sources resulted in more than a thousand unique entries. After manually removing all obvious noneligible companies, a final list of 592 high-tech companies was then put together.

To maximize our response rate, we contracted Léger360 (now Léger), a well-known survey company, to administer the questionnaire and find new companies eligible for our study. Their professional approach is appreciated by firms that have responded to surveys in the past. Data collection began in September 2016 based on a convenience sampling method. To cover even more ground, we also used the snowball method. More precisely, respondents first had to answer the eligibility question: "Does your company develop or/and commercialize nanotechnologies and/or advanced materials?" Then, if the firms were eligible, the respondents were asked whether or not they were interested in participating in our study. Finally, they were asked whether they had any recommendations for any potential respondents who might be eligible for our investigation as an attempt to increase our sample size. Firms that agreed to participate in the survey but did not complete the questionnaire received two telephone reminders at 1-week intervals. After 332 solicitations, the incidence rate was 28%, which indicated that 93 companies were eligible.

To maximize our sample size, we added companies to the list by using the common characteristics of the eligible respondents. Companies that were eligible for the study (the 93 eligible companies listed above) were grouped and classified according to their North American Industry Classification System (NAICS) codes. Twenty-three six-digit NAICS codes corresponding to 67% of the eligible companies were identified. Léger360 acquired a list of 3,345 companies representative of the frequency distribution of the NAICS code obtained from InfoGroup.

We solicited 2,971 Canadian high-tech companies through the entire data acquisition process. Of these, 973 companies did not respond, 1,459 were not eligible, 380 refused to participate, and 168 eligible companies agreed to participate, of which 89 respondents completed the survey. Because of Léger360's code of ethics, we have no means of identifying whether respondents refused to participate before or after answering the eligibility questions. The unknown nature of the nonrespondents necessitates an approximation of the response rate. Assuming a uniform distribution, a 12% response rate was obtained by assigning the same characteristics to respondents who answered the eligibility questions to nonrespondents and respondents who refused to participate (see Table 1). As the population is unknown, we have a nonprobabilistic convenience sample for which the methodology may have induced a selection bias. In addition, we assume that the respondents were honest and answered the survey with goodwill.

The resulting sample represents a diverse selection of Canadian high-tech firms that develop or use nanotechnologies or advanced materials. Moreover, 74% of the firms are considered nanotechnology or advanced materials intensive, which means that at least 80% of their revenues come from nanotechnology or advanced materials innovations. The different application domains

Table 1. Estimated response rate

Number of high technological Canadian companies:	
Reached (A)	2,971
Did not answer (B)	973
Refused to participate (C)	380
Unknown status (A − B − C = D)	1,618
Not eligible (E)	1,459
[Rate of Not Eligible (E/A = F)]	[49%]
Estimated not eligible (D × F = G)	795
Estimated eligible (A − E − G = H)	718
Total accepted	168
Total completed	89
[Estimated response rate]	[12%]

are as follows: 54% in advanced materials, 21% in biotechnology and medicine, 24.4% in electronics, 23.3% in equipment and devices, 13.3% in photonics, and 33.3% in other domains. More than 50% of respondents are small businesses and 83.5% are SMEs. The $94 million revenue average decreases to $31 million when the three largest firms are excluded. Finally, 85% of the firms come from Quebec (54.5%) and Ontario (30.7%), and 12% are from British Columbia and Alberta.

To test several types of biases, such as self-reporting bias and nonrespondent bias, we selected 79 eligible enterprises that did not participate in the study as a control sample. The Canadian government's[1] Directory of Canadian Companies was used as an external source of data to rule out the possibility of nonresponse bias. The database of companies from different sectors comprises information provided by the companies themselves on a voluntary basis. Although the Canadian government does not guarantee the accuracy or the reliability of the content, we assumed that the companies that willingly update information in an official public database input accurate information. We therefore assume that this mitigates the self-reporting bias that might arise from this type of source. The database provides the number of employees for 37 firms and revenues for 30 firms from our main sample, as well as the number of employees for 29 firms and revenues for 26 firms from our control sample. Using a two-tailed Mann-Whitney U test, we compared the main sample with the control group for these two metrics. We did not find any significant difference (at the 5% level) between the two samples for either metric ($p = 0.115$; $p = 0.166$ for the number of employees and revenues respectively), which suggests the absence of a nonresponse bias: There are no significant differences between the characteristics of the respondents and the nonrespondents from our list and thus the sample of eligible companies is homogeneous.

We then compared the data obtained from our questionnaire-based survey with the data from the Directory of Canadian Companies using the same two metrics to verify whether an important self-reporting bias exists. For every firm for which we had data from both our questionnaire-based

[1] https://www.canada.ca/en/services/business/research/directoriescanadiancompanies.html, accessed June 4, 2020.

survey and from the Directory, we tested each data pair with a two-tailed Wilcoxon Signed Ranks Test. Once again, we did not find any significant difference (at the 5% level) between our questionnaire-based survey results and the data from the Directory ($p = 0.058$; $p = 0.714$), which suggests that the self-reporting bias issuing from the questionnaire-based survey is no different than that found in an official public database.

To build the four factors described above, R&D, IP protection, collaboration, and external financing, we identified all the relevant questions from the questionnaire-based survey and transformed the answers to these questions into different types of variables. The 12 questions used are listed in Appendix 1. Moreover, we transformed every continuous variable from the survey that did not follow a normal distribution by applying a natural logarithm (ln) or an inverse function (inv). Some Likert-scale questions could not be normalized because they were skewed on one tail or the other. We thus transformed them into dummy variables by attributing the value of 1 to answers associated with important (5), very important (6), and essential (7) and the remaining were given the value of 0.

A principal component analysis (PCA) was then performed to validate reflective constructs. Only constructs with sufficiently high Kaiser-Meyer-Olkin (KMO) measures (KMO > 0.6) were deemed valid, and only dimensions with Cronbach's alpha above 0.7 (Hair et al., 1998) were considered reliable. We used PCA with a Varimax rotation on seven-point Likert-scale questions that described the concept of R&D (R&D questions 2, 3, 5, and 6 in Appendix 1) to explore whether any particular combination of variables could lead to relevant reflective dimensions corresponding to specific factors of the concept examined. Two factors were created, but neither the KMO measure nor Cronbach's alpha reached an acceptable level (KMO < 0.6, alpha < 0.7) to satisfy the validity and reliability of the constructs. In addition, these combined variables did not correlate with each other, which suggested using a formative construct. We thus proceeded to treat each item individually rather than as a composite indicator.

In the end, we generated nine variables pertinent to R&D, one variable relating to collaboration, two variables corresponding to external financing, and two variables measuring IP. Table 2 describes the details of the questionnaire-based survey constructs. Descriptive statistics are presented in Appendix 2.

3.2. Web-Based Data Collection

While we were processing the questionnaire-based survey data, we initialized the indexing robot with Nutch[2], an open source Web scraper software. We provided the robot with an initial list of uniform resource locators (URL) corresponding to the 89 enterprises that compose our questionnaire-based data set. This list is the first level of indexing traversal by the robot, which then browses the structure of the entire website to extract all the URLs. To begin the data collection process, the URLs linked to a company's website are written in a text file. The URLs are then injected into Nutch's database.

Once the database has been populated with the list of sites to be visited, the robot's route is generated by indicating the maximum number of links per page that we want to collect. The robot ranks the pages with a score that it calculates to prioritize the text collection. For instance, a page will have a high score if it is pointed to by many other pages. In addition, the higher the scores of pages pointing to another page, the higher the score of the page pointed to. Nutch then selects the links of the pages with the highest score to be collected first. When the links to browse are

[2] We used Apache Nutch 2.2.1 version. Documentation can be found on https://nutch.apache.org/apidocs/apidocs-2.2.1/index.html, accessed June 4, 2020.

Table 2. Questionnaire-based survey construct

Concepts	Indicators	Variables
R&D	Number of R&D projects in nanotechnology/advanced materials	ln(nb_R&D_proj) (Continuous)
	Dichotomized importance of internal R&D as a source of knowledge	dInt_R&D_source (Dummy)
	Level of importance of Commercial laboratories/R&D firms/Technical Consultants as a source of knowledge	Ext_R&D_source (Ordinal)
	Level of importance of contracting of external R&D service providers	Contract_In_R&D (Ordinal)
	Level of importance of providing R&D services to third parties	Contract_Out_R&D (Ordinal)
	Time of R&D	ln(time_R&D) (Continuous)
	Dichotomized importance of Private research laboratories/Research and Development firms as collaborators for the development and the commercialization	dCollab_R&D (Dummy)
	Dichotomized importance of accessing research and development from collaborators for the development and the commercialization	dCollab_R&D_Reason (Dummy)
	Proportion of Canadian employees assigned primarily in R&D (%)	propEmp_R&D (Continuous)
IP	Number of IP mechanisms used	nbIP (Count data)
	Number of patents	ln(nbPatent) (Continuous)
Collaboration	Use of collaboration for the latest innovation	dCollab (Dummy)
External financing	Proportion of external financing for R&D (%)	PropExtFin_R&D (Continuous)
	Proportion of external financing for commercialization (%)	PropExtFin_Comm (Continuous)
	Total proportion of external financing (%)	propTotExtFin (Continuous)

Note: All continuous and ordinal variables are normal.

generated, Nutch will collect all the HTML code from the pages as well as the links to the other pages.

There are mainly three different types of links that structure the Internet network. Upstream links start from a page to one or more pages not necessarily on the same site. Downstream links are all links pointing to the same page. Finally, vertical links are all links within the same website (Van de Lei & Cunningham, 2006). The indexing robot scans all links without distinguishing between domain names and ranks them by score to first scan the links it considers most interesting. We have therefore limited the robot to the vertical links of the domain names present in the initial list. This is possible by modifying the Nutch configuration file limiting the robot to the initial list of domains.

By default, Nutch indexes only the text of pages in hypertext markup language (HTML) format. However, relevant information was also stored in other digital formats. We used the Tika plug-in to browse and index the text of documents in portable document format files (.pdf) and Microsoft Word documents (.doc and .docx) format.

A further difficulty stems from the fact that our sample included companies that use multiple languages on their websites. Most Canadian companies will have English or French content and some will have other languages. Therefore, we did take into account the fact that the information

could be given in languages other than English. There were several different cases to deal with. If a website was entirely in English, which corresponds to a majority of Canadian company websites, we then kept all the data from the website. If a website was available in several languages, including English, then only pages written in English were taken into account. If the website was exclusively available in French, the text was translated into English. Finally, the last case was a page whose language was neither French nor English. Such a page was not considered eligible for our study and, therefore, was not kept for the rest of the study.

Language detection was performed by the Compact Language Detector software and the translation was performed with Google's programming interface, Goslate. Translation with Goslate was also used by Arora, Youtie, et al. (2013) to analyze the content of Chinese graphene-producing company websites.

We assigned to each recovered page a label of language, "French," "English," or "other" and translated the French pages into English for the relevant sites according to our assumptions.

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. In the end, more than 9.7 million words from 27,000 pages were extracted from the complete texts of corporate websites. We then used a word frequency analysis on the text present in the websites. More specifically, for the 79 websites captured, we gathered information on the four innovation and commercialization factors mentioned above. For each factor, we listed all the relevant keywords that appear in corporate websites. Figure 1 shows the complete process to build the web-based innovation indicators to be compared with the questionnaire-based indicators described in the previous section.

Factors, keywords, and the web mining constructs are described in Table 3. R&D (Gök et al., 2015) and collaboration (Ramdani, 2014) keywords were selected from the literature, and IP and external financing were identified from our own investigations of the literature. The most relevant keywords of any paper, which are generally listed on its first page, particularly in the abstract[3], serve as a basis for the list of keywords used for the construction of our factors. We used additional keywords related to specific public funds and programs. As mentioned earlier, Canada's public programs and funding opportunities for companies for the development of nanotechnology and advanced materials projects are listed on the ISED website. Specific information on these funds and programs was converted into keywords tied to the external funding factor for the individual firms for which we extracted website data. Keywords were manually tested to ensure they did not lead to false positive results.

Clustering using keyword frequency analysis with RapidMiner, a text mining program, enabled us to count the number of occurrences of each keyword for each factor. We transformed these clusters of occurrences into four continuous variables. Because the 79 companies differ in structure and size, and therefore present different quantities of information on 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 1,000. For each continuous variable, we calculated the kurtosis and skewness measures to determine whether they followed a normal distribution. Given that none of the four variables did so, they were transformed by applying a natural logarithm (ln) or an inverse function (inv). In the case of external financing, because we did not reach normality with either transformation, this variable was therefore treated using nonparametric tests that do not require normality.

[3] For this pilot/exploratory study, we did not mine entire articles to select the most common or important keywords related to specific contexts.

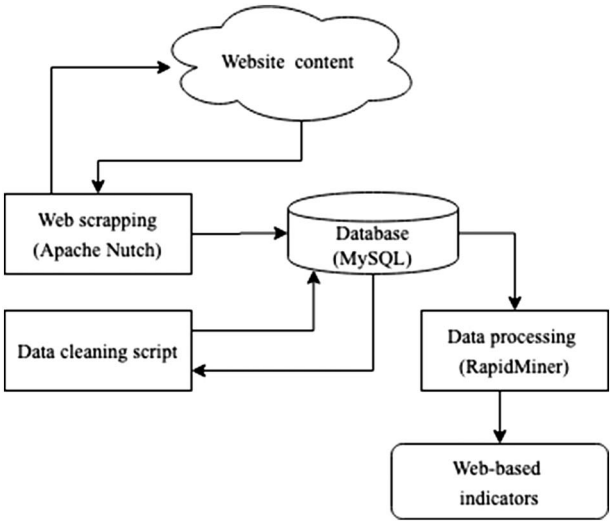


Figure 1. Data mining and treatment process.

A possible selection bias arises from the fact that we selected only those companies that answered the survey. To verify whether this was the case, we ran our web mining program on the websites of the control sample and generated the same variables. We then used a two-tailed Student's t test to test the difference in means for the following variables: ln(WEB_R&D)

Table 3. The web mining construct

Factors	Keywords	Indicators	Variables
R&D	research & development, research and development, r&d, researcher, product development, technology development, technical development, development phase, development program, development process, development project, development cent, development facility, technological development, development effort, development cycle, development research, development activity, fundamental research, basic research (Gök et al., 2015)	Natural logarithm of the number of keywords frequencies divided by the total number of words multiplied by 1,000	ln(WEB_R&D)
IP	patent, intellectual property, trade secret, industrial design	Inverse (1/x) of keywords frequencies divided by the total number of words multiplied by 1,000	inv(WEB_IP)
Collaboration	affiliation, collaboration, cooperation, partners, partnership, consorti, international consorti, global consorti	Natural logarithm of the number of keywords frequencies divided by the total number of words multiplied by 1,000	ln(WEB_Collab)
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	The number of keywords frequencies divided by the total number of words multiplied by 1,000	WEB_ExtFin*

* All transformed variables are normal except WEB_ExtFin, which could not be normalized by any transformation.

($p = 0.130$), inv(WEB_IP) ($p = 0.083$) and ln(WEB_Collab) ($p = 0.144$), and concluded that there is a nonsignificant (at 5% level) difference between the two samples for these three variables. Finally, using a two-tailed Mann-Whitney U test, we compared the means of the variable WEB_ExtFin ($p = 0.008$) from both our sample and the control sample and found a significant difference. For this particular variable, we cannot conclude that the means of the two samples are the same. A selection bias is therefore present for the variable WEB_ExtFin and will be included in the limitations of our research.

3.3. The Multitrait Multimethod (MTMM) Matrix Technique

The literature describes two types of biases that can threaten construct validation: (a) Mono-method biases can occur when the by-products from the unique combination of a trait (in our case an innovation factor such as R&D, IP, collaboration, and external financing) and a measure introduce a systematic variance; and (b) by-products that are congeneric (i.e., inherent in the methods used) generate anomalies among the items forming the measures (Ortiz de Guinea, Titah, & Léger, 2013; Reinig, Briggs, & Nunamaker, 2007; Straub & Burton-Jones, 2007; Straub, Limayem, & Karahanna-Evaristo, 1995).

First introduced by Campbell and Fiske (1959), the MTMM is designed for the convergent and discriminant validation of a construct where a set of *t* traits (interchangeable with factors in our case) are measured with *m* different methods. It has been suggested that this method is an effective way to verify the ability of new measurement methods to successfully measure what they are supposed to, while testing for the presence of mono-method biases in social science studies, such as psychology (Bar-Anan & Vianello, 2018; Guo, Aveyard, et al., 2008), education (Campbell, Michel, et al., 2019; Gulek, 1999), organizational research (Bagozzi, Yi, & Phillips, 1991), marketing research (Lugtig, 2017), and information systems (Ortiz de Guinea et al., 2013). Even though its use for construct validity has been employed selectively since its introduction in 1959, the MTMM matrix is still one of the very few possible ways to attempt construct validation and this is why we have been witnessing a resurgence in its use in the past few years. To our knowledge, this validation method has not been used in innovation studies or to assess new innovation metrics compared to the traditional means of measuring innovation, as in the Oslo Manual (OECD & Statistical Office of the European Communities, 2005; OECD & Eurostat, 2019).

An example of an MTMM matrix with three traits (A, B, C) measured using three different methods is shown in Table 4. The matrix easily allows the visualization of four different means of assessing the quality of a measure: reliability, convergent validation, absence of either a combination of trait and method effects or a mono-method bias and, finally, discriminant validation (Campbell & Fiske, 1959).

The diagonal of the matrix (also called the *reliability diagonal*) comprises six Cronbach alpha values (α_{A1}, α_{B1}, α_{C1}, α_{A2}, α_{B2}, and α_{C2}), which measure the correlation of the traits with themselves and are thus an indication of the reliability of each measure for each trait (Campbell & Fiske, 1959). These mono-trait mono-method values indicate the reliability of a reflective measure (Campbell & Fiske, 1959), which is required to ensure that a measure will produce the same result under the same conditions. High Cronbach's alpha values on the matrix diagonal confirm the reliability of the traits; that is, that the variance of the measures behaves systematically (Churchill, 1979). In other words, any possible occurrence of a random error is minimized.

The diagonal at the cross-section of Method 1 and Method 2 (also called the *mono-trait hetero-method*, or *validity diagonal*) is comprised of three correlation values ($r_{A1,A2}$, $r_{B1,B2}$, and $r_{C1,C2}$) that compare one trait measured with two different methods. High and significant values confirm

Table 4. The MTMM matrix technique

		Method 1			Method 2		
	Traits	A1	B1	C2	A2	B2	C2
Method 1	A1	$\propto_{A1}$					
	B1	$r_{A1,B1}$	$\propto_{B1}$				
	C1	$r_{A1,C1}$	$r_{B1,C1}$	$\propto_{C1}$			
Method 2	A2	$r_{A1,A2}$	$r_{B1,A2}$	$r_{C2,A2}$	$\propto_{A2}$		
	B2	$r_{A1,B2}$	$r_{B1,B2}$	$r_{C2,B2}$	$r_{A2,B2}$	$\propto_{B2}$	
	C2	$r_{A1,C2}$	$r_{B1,C2}$	$r_{C2,C2}$	$r_{A2,C2}$	$r_{B2,C2}$	$\propto_{C2}$

the *convergent validity* of the methods (Campbell & Fiske, 1959); that is, that two different methods measure the same traits (Churchill, 1979).

Two triangles, called *hetero-trait mono-method triangles*, represent the cross-sections of traits that belong to the same method within the matrix. They are comprised of three correlation values per triangle. For instance, the hetero-trait mono-method triangle of Method 1 comprises $r_{A1,B1}$, $r_{A1,C1}$, and $r_{B1,C1}$, and that of Method 2 comprises $r_{A2,B2}$, $r_{A2,C2}$, and $r_{B2,C2}$ (Campbell & Fiske, 1959). High correlation values (higher than the validity diagonal) within the hetero-trait mono-method triangle question the validity of the construct because of either a combination of trait and method effects or a mono-method bias.

Finally, two other triangles that represent the cross-sections of traits that belong to different methods within the matrix (also called *hetero-trait hetero-method triangles*) are comprised of three correlation values per triangle. The hetero-trait hetero-method triangle of Method 1 is comprised of $r_{A1,B2}$, $r_{A1,C2}$, and $r_{B1,C2}$, and that of Method 2 is comprised of $r_{B1,A2}$, $r_{C2,A2}$, and $r_{C2,B2}$ (Campbell & Fiske, 1959). Again, high values (higher than the validity diagonal) question the *discriminant validity* of the methods. The discriminant validity is used to assess whether a method fails to measure something it is not supposed to measure effectively (Peter & Churchill, 1986).

Bagozzi et al. (1991) have raised questions about the reasonability of the criteria needed, the lack of standard practice, the negligence of the amplitude of the differences between pairs of correlations in the analysis, and the lack of information about the nature of the variation in measures due to traits, method or random error. Fiske and Campbell (1992) listed the exceptional conditions required to successfully benefit from the method. Bagozzi et al. (1991) suggested CFA as an alternative to the MTMM for construct validation, which, although it does not share the same limitations, is unable to differentiate the random error from the variance measure and unable to verify interactions between traits and methods. Therefore, performing both methods will mitigate the limitations inherent in either method taken individually and ensure robustness of the results. In this paper, we adopt the methodology used by Ortiz de Guinea et al. (2013), who opted for an MTMM analysis followed by a post hoc analysis with a CFA to validate their results.

4. RESULTS

4.1. Keywords Analysis

More than 9.7 million words from 27,000 pages were extracted from the complete texts of 79 corporate websites. The details of the frequency analysis for each indicator are provided in

Table 5. For the R&D factor, 1,974 keywords were counted, from which six keywords (r&d, scientist, laboratory, research and development, researcher, laboratorie) account for more than 85% of the total frequency distribution. For IP, 75% of the total frequency distribution (397) is embodied by the keyword "patent," hence suggesting that most of the signal provided by the IP indicator is actually related to patenting. This is hardly surprising considering our sample of nanotech firms. For collaboration of 1,769 keywords were counted, from which

Table 5. Keyword frequencies

R&D		IP		Collaboration		External financing	
r&d	480	Patent	300	partner	967	grant	202
scientist	368	intellectual propert	84	partnership	422	venture capital	31
laboratory	274	trade secret	8	collaboration	132	business development bank of canada	7
research and development	246	patent pending	4	alliance	88	industrial research assistance program	4
researcher	185	patent protect	1	collaborative	62	private investment	4
laboratorie	165			cooperation	35	atlantic canada opportunitie	1
product development	99			partnering	34	public fund	1
technology development	64			collaborating	18	public funding	1
research & development	33			cooperative	9		
development phase	15			cooperating	2		
development project	12						
development effort	10						
basic research	5						
technological development	4						
development program	4						
fundamental research	2						
development cycle	2						
development center	2						
development research	1						
development process	1						
development facility	1						
development faciliti	1						
Total	1,974	397		1,769		251	
Mean	89.72	66.17		147.41		31.38	
Std	138.23	118.91		283.20		69.67	

more than half are associated with partnering activities. On external financing, we only obtained 251 keywords, from which, more than 80% of the frequency distribution was dominated by the word "grant."

4.2. Testing Proposition 1

Construct validation

Each pair of variables related to the same concept from the two methods (web mining and questionnaire-based survey) was compared, using a Pearson correlation analysis when the variables followed a normal distribution, and a Spearman correlation when they did not, to assess whether the variables stemming from the web mining technique could be used as a proxy for similar concepts measured by a survey. Details of the construct comparison of the web-based indicators and the questionnaire-based indicators are provided in Table 6.

Correlation results

All correlation results between the web-based indicators and the questionnaire-based indicators are detailed in Appendix 3. For this study, the correlations were tested with two-tailed test at the 5% level of significance. Our results show a correlation of 0.306 ($p < 0.01$) between the R&D web-based indicator and whether a firm is likely or not to provide R&D services to third parties (from the questionnaire). Additionally, we find a correlation of 0.306 ($p < 0.01$) when we associate the R&D web-based indicator with whether or not a firm has a high percentage of employees working on R&D tasks. Moreover, a correlation of 0.284 ($p < 0.05$) is observed

Table 6. Validation construct for testing proposition 1

Concepts	Web-based variables	Questionnaire variables
R&D	ln(WEB_R&D)	ln(nb_R&D_proj)
		dInt_R&D_source
		Ext_R&D_source
		Contract_In_R&D
		Contract_Out_R&D
		ln(time_R&D)
		propEmp_R&D
		dCollab_R&D
		dCollab_R&D_Reason
IP	Inv(WEB_IP)	nbIP
		ln(nbPatent)
Collaboration	ln(WEB_Collab)	dCollab
External financing	WEB_ExtFin	propExtFin_R&D
		propExtFin_Comm
		propTotExtFin

between the R&D web-based indicator and whether or not a firm is likely to contract R&D services from external providers. Finally, we find a nonsignificant correlation of 0.197 (p = 0.100) when we associate the R&D web-based indicator with whether or not a firm has a long R&D process. The fact that the variable related to the number of R&D projects is not correlated with our R&D web-based indicator (r = 0.002; $p > 0.1$) strongly suggests that the latter does not properly account for the size dimension of R&D activities.

The IP web-based indicator strongly correlates with the variables from the survey regarding the use of IP protection mechanisms (r = 0.368; $p < 0.01$) and patent-related activities (r = 0.396; p = 0.5). The web content analysis method seems to be able to appropriately capture the importance of the use of IP mechanisms using our sample.

Relations between the web-mining and questionnaire-based indicators for the other two factors are not as strong. For instance, the collaboration web-based indicator is weakly correlated and nonsignificant at 5% (r = 0.222; $p < 0.1$) with the questionnaire-based indicator for the firms that are confirmed as having collaborated. Finally, the external financing web-based indicator is also weakly correlated and nonsignificant at 5% with the extent of the use of external funding for commercialization purposes (r = 0.222; $p < 0.1$).

Consequently, we cannot reach a definite conclusion regarding proposition 1, especially in the case of our web-based indicators for collaboration and external financing (a less constraining significance level of 10% would be required to accept the correlations). In the following paragraphs, we will restrict our analyses to the R&D and IP factors to test propositions 1 and 2.

MTMM results

All the variables used to measure the R&D and IP factors from the questionnaire-based survey that correlated significantly with our web-based indicators were selected. As explained in section 3.3, the MTMM compares different combinations of methods that measure the same traits (factors). Thus, all possible combinations of traits and methods need to be tested. In the case of the web-based indicators, only one combination of trait and method is used. For the questionnaire-based indicators, three different variables are correlated with the R&D web-based indicator and two different variables are correlated with the IP web-based indicator. As such, six different combinations are possible and thus six different matrices are built. To ease the reader's comprehension of the interpretation of an MTMM matrix, Table 7 shows an example of how the results are displayed, along with a short analysis of all the different validity tests below the table. Furthermore, all traits are considered to be single items, which implies that we cannot calculate the reliability of the measure. Accordingly, the reliability diagonal will be neglected in our analysis.

In all matrices, we can observe that the hetero-trait mono-method value is low and not significant for the web-based method ($r_{RD1,IP1} = -0.180$; p-value > 0.05) and lower than all corresponding mono-trait hetero-method diagonal values ($r_{RD1,RD2}$ and $r_{IP1,IP2}$). This suggests that there is no combination of trait and method effects and that no method bias for the web-based method is present. The complete analysis of all the different matrices results is shown in Tables 8–13.

Of the six MTMM matrices produced, four matrices yield acceptable results:

- Level of importance of contracting R&D with the number of patents (Table 9)
- Level of importance of providing R&D with the number of IP mechanisms used (Table 10)
- Level of importance of providing R&D with the number of patents (Table 11)
- Proportion of employees assigned primarily in R&D with the number of IP mechanisms used (although it is a weak construct) (Table 12)

Table 7. MTMM matrix output example

		Web method		Questionnaire method	
	Traits	RD1	IP1	RD2	IP2
Web method	RD1	–[a]			
	IP1	$r_{RD1,IP1}$	–[a]		
Questionnaire method	RD2	$r_{RD1,RD2}$	$r_{IP1,RD2}$	–[a]	
	IP2	$r_{RD1,IP2}$	$r_{IP1,IP2}$	$r_{RD2,IP2}$	–[a]

Convergence validity: **Yes** if $r_{RD1,RD2}$ and $r_{IP1,IP2}$ are high and significant;

Discriminant validity: **Yes** if $r_{RD1,RD2}$ and $r_{IP1,IP2}$ are both[b] > $|r_{RD1,IP2}|$ and $|r_{IP1,RD2}|$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes** if $r_{RD1,RD2}$ and $r_{IP2,IP2} > |r_{IP1,RD1}|$ and $|r_{RD2,IP2}|$.

Notes (common to Tables 8–13):

[a] All traits are measured by single items, no reliability statistic can be calculated.

[b] *To increase the readability of the results in each table what we mean by "are both" is:* $r_{RD1,RD2} > |r_{RD1,IP2}|$, $r_{RD1,RD2} > |r_{IP1,RD2}|$, $r_{IP2,IP2} > |r_{RD1,IP2}|$, and $r_{IP2,IP2} > |r_{IP1,RD2}|$

* $p < 0.05$; ** $p < 0.01$

The rejection of the validity of the level of importance of contracting R&D with the number of IP mechanisms used construct (Table 8) suggests the presence of either a combination of trait and method effects or a mono-method bias, as the correlation between the two items from the questionnaire is stronger than the correlation obtained with their corresponding web-based measure. Thus, we cannot effectively discriminate the two web-based indicators from the questionnaire-based indicators, which implies that they cannot be part of the same construct.

Table 8. Level of importance of contracting R&D with the number of IP mechanisms used

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	Contract_In_R&D	nbIP
Web method	WEB_RD	–[a]			
	WEB_IP	–0.18	–[a]		
Questionnaire method	Contract_In_R&D	0.283*	0.209	–[a]	
	nbIP	0.007	0.368**	0.433**	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.283$* and $r_{IP1,IP2} = 0.368$** are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.283$* and $r_{IP1,IP2} = 0.368$** >> $|r_{RD1,IP2}| = 0.007$ and $|r_{IP1,RD2}| = 0.209$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **No**, $|r_{RD2,IP2}| = 0.433$** >> $r_{RD1,RD2} = 0.283$* and $r_{IP1,IP2} = 0.368$**.

Table 9. Level of importance of contracting R&D with the number of patents

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	Contract_In_R&D	ln(nbPatent)
Web method	WEB_RD	–[a]			
	WEB_IP	−0.18	–[a]		
Questionnaire method	Contract_In_R&D	0.283*	0.209	–[a]	
	ln(nbPatent)	0.044	0.395*	−0.077	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.283^*$ and $r_{IP1,IP2} = 0.395^*$ are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.283^*$ and $r_{IP1,IP2} = 0.395^* >> |r_{RD1,IP2}| = 0.044$ and $|r_{IP1,RD2}| = 0.209$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes**, $r_{RD1,RD2} = 0.283^*$ and $r_{IP1,IP2} = 0.395^* >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.077$.

The rejection of the discriminant validity between the web-mining method and the questionnaire-based survey method with the proportion of employees assigned primarily in R&D with the number of patents (Table 13) stems from the fact that the correlation is higher with the IP web-based indicator than with the R&D web-based indicator. After all, it can be expected that a high proportion of employees dedicated at R&D activities should correlate strongly with the number of patents. Therefore, these indicators do not discriminate each other, and this construct is thus rejected.

Finally, the proportion of employees assigned primarily in R&D with number of IP mechanisms used passes the validity test because the four correlations are lower than the corresponding values found in the validity diagonal, which is the accepting criterion even if $|r_{IP1,RD2}| = 0.288^*$

Table 10. Level of importance of providing R&D with the number of IP mechanisms used

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	Contract_Out_R&D	nbIP
Web method	WEB_RD	–[a]			
	WEB_IP	−0.18	–[a]		
Questionnaire method	Contract_Out_R&D	0.306*	0.068	–[a]	
	nbIP	0.007	0.368**	−0.066	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^*$ and $r_{IP1,IP2} = 0.368^{**}$ are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.306^*$ and $r_{IP1,IP2} = 0.368^{**} >> |r_{RD1,IP2}| = 0.007$ and $|r_{IP1,RD2}| = 0.068$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes**, $r_{RD1,RD2} = 0.306^*$ and $r_{IP1,IP2} = 0.368^{**} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.066$.

Table 11. Level of importance of providing R&D with the number of patents

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	Contract_Out_R&D	ln(nbPatent)
Web method	WEB_RD	–[a]			
	WEB_IP	–0.18	–[a]		
Questionnaire method	Contract_Out_R&D	0.306**	0.068	–[a]	
	ln(nbPatent)	0.044	0.396*	–0.134	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*}$ are high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} >> |r_{RD1,IP2}| = 0.044$ and $|r_{IP1,RD2}| = 0.068$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.134$.

and $|r_{RD1,IP2}| = 0.284^{*}$ are reasonably high and significant. However, it is determined to be a weak construct because of the slight differences in correlation.

Although these results are encouraging, the presence of four valid constructs ironically shows that we cannot isolate precise actions with our web-based indicators (see the results summary in Table 14). In other words, the "real" meaning of these indicators remains unknown and, as a result, proposition 1 is not confirmed. However, these measures may need to be put into a broader perspective to be effective, which may provide a better fit for the scope allowed by proposition 2. The next steps in this paper test whether the web-based indicators represent factors that are considered important for the firms.

Table 12. Proportion of employees assigned primarily in R&D with the number of IP mechanisms used

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	propEmp_R&D	nbIP
Web method	WEB_RD	–[a]			
	WEB_IP	–0.18	–[a]		
Questionnaire method	propEmp_R&D	0.306**	0.288*	–[a]	
	nbIP	0.007	0.368**	0.284*	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.368^{**}$ are high and significant;

Discriminant validity: **Weak yes**, $r_{RD1,RD2} = 0.306^{*} >> |r_{RD1,IP2}| = 0.007$, but $r_{RD1,RD2} = 0.306^{*} > |r_{IP1,RD2}| = 0.288^{*}$ which is also high and significant; finally, $r_{IP1,IP2} = 0.368^{**} >> |r_{IP1,RD2}| = 0.288^{*}$ and $|r_{RD1,IP2}| = 0.007$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Weak yes**, $r_{RD1,RD2} = 0.306^{*} >> |r_{IP1,RD1}| = 0.18$, but $r_{RD1,RD2} = 0.306^{*} > |r_{RD2,IP2}| = 0.284^{*}$ which is also high and significant; finally, $r_{IP1,IP2} = 0.368^{**} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD1,IP2}| = 0.284^{*}$.

Table 13. Proportion of employees assigned primarily in R&D with the number of patents

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	propEmp_R&D	ln(nbPatent)
Web method	WEB_RD	–[a]			
	WEB_IP	−0.18	–[a]		
Questionnaire method	propEmp_R&D	0.306**	0.455**	–[a]	
	ln(nbPatent)	0.044	0.396*	0.02	–[a]

Convergence validity: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*}$ are high and significant;

Discriminant validity: **No**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} << |r_{IP1,RD2}| = 0.455^{**}$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes**, $r_{RD1,RD2} = 0.306^{**}$ and $r_{IP1,IP2} = 0.396^{*} >> |r_{IP1,RD1}| = 0.18$ and $|r_{RD2,IP2}| = 0.02$.

4.3. Testing Proposition 2

Construction of formative indices and new validation construct

Given the vast number of words used to construct the web-based indicators, as we concluded in the previous section, treating each questionnaire-based variable individually may not be appropriate. To illustrate the large lexical field of possible words related to the factors studied, and to properly test proposition 2, it is conceptually sound to build one single measure: one formative

Table 14. Summary of construct validity results

Table		Convergent validity	Discriminant validity	Construct validity without either a combination of trait and method effects or a mono-method bias
8	Level of importance of contracting R&D with the number of IP mechanisms used	Yes	Yes	No
9	Level of importance of contracting R&D with the number of patents	Yes	Yes	Yes
10	Level of importance of providing R&D with the number of IP mechanisms used	Yes	Yes	Yes
11	Level of importance of providing R&D with the number of patents	Yes	Yes	Yes
12	Prop. of employees assigned primarily in R&D with the number of IP mechanisms used	Yes	Weak yes	Weak yes
13	Prop. of employees assigned primarily in R&D with the number of patents	Yes	No	Yes

index with all the questions related to R&D and IP. Because the PCA performed on all the items related to R&D and to IP presented at the end of section 3.1 did not produce any significant KMO and Cronbach's alpha measures, the use of a formative index comprising several subelements explaining our R&D factor in its broadest sense may be more appropriate.

Partial least square (PLS) regressions were estimated to determine whether it is possible to create valid formative indices for R&D and IP. To use PLS regressions, the methodology requires only the use of complete data sets (Nelson, Taylor, & MacGregor, 1996). Nonresponse is usually treated by either weight adjustment (i.e., deleting incomplete data entry and weighting remaining respondents to compensate for the deletion) or imputation (i.e., adding artificial values based on average by classes and editing methods (Särndal, Swensson, & Wretman, 1992) to replace the missing values) (Haziza & Beaumont, 2007). As our sample size for IP is already low for one of the items (39 for the number of patents), we could not afford to treat the missing data with a weight adjustment. Accordingly, we replaced the missing data with their imputation class based on control variables. We sorted the firms by sector, then by number of employees, and then by revenue. Depending on the situation, we used the mean of the class or the most conservative nearest-neighbor, a method commonly used in the literature (Haziza & Beaumont, 2007; Little, 1986; Thomsen, 1973).

Because not all the items shared the same scale, we transformed each variable into a Z-score. PLS regressions were then estimated using WarpPLS 5.0 software with the following settings: MODEL B BASIC Warp3 Stable 3 and MODEL B BASIC Linear Stable 3. The two different settings produced the same conclusions. The details of the construct comparing the web mining technique and the questionnaire-based survey are shown in Table 15 (the PLS regressions and the resulting weights used to build these indices are provided in Appendix 4).

All weights are significant ($p < 0.01$); indicator weight-loading signs are all positive; variance inflation factors (VIF) are all very low (< 1.5); and the effect sizes (ES) are all greater than 0.02. All the criteria are met to indicate that the indices generated are valid (Cenfetelli & Bassellier, 2009; Cohen, 1988; Diamantopoulos, 1999; Diamantopoulos & Siguaw, 2006; Diamantopoulos & Winklhofer, 2001; Petter, Straub, & Rai, 2007). For each factor, the sum of each weighted variable generated both indicators R&D_INDEX and IP_INDEX.

Table 15. The validation construct for proposition 2 testing

Concepts	Web mining variables	PLS-built indices	Questionnaire variables
R&D	ln(WEB_R&D)	R&D_INDEX	Z_nb_R&D_proj
			Z_Int_R&D_source
			Z_Ext_R&D_source
			Z_Contract_In_R&D
			Z_Contract_Out_R&D
			Z_time_R&D
			Z_propEmp_R&D
IP	Inv(WEB_IP)	IP_INDEX	Z_nbIP
			Z_nbPatent

Note: All variables are continuous and normal.

Final MTMM analysis

This final MTMM matrix includes the two web-based indicators along with the R&D_INDEX and IP_INDEX (see Table 16). Once again, the reliability diagonal will be neglected in our analysis, as the measures are made with single items from the web method and with formative indices from our questionnaire. The mono-trait hetero-method diagonal shows high and significant correlations for R&D ($r = 0.419$; $p < 0.01$) and for IP ($r = 0.520$; $p < 0.01$), which hints at strong convergent validity. The hetero-trait mono-method value is low and not significant for the web content analysis method ($r = 0.182$; $p > 0.05$), although the questionnaire-based survey method value is high and significant ($r = 0.320$; $p < 0.01$). However, the mono-trait hetero-method values ($r = 0.419$ and $r = 0.520$) for R&D and IP respectively are much higher than the hetero-trait mono-method values ($r = 0.320$ and $r = 0.182$), which indicates no combination of trait and method effects and no mono-method biases. The first hetero-trait hetero-method value is low and not significant ($r = -0.017$; $p > 0.05$), and the other is moderate and significant ($r = 0.294$; $p < 0.05$). However, and more importantly, the correlations are much lower than the corresponding values found in the validity diagonal, which shows good discriminant validity. As all the conditions are satisfied under the original guidelines proposed by Campbell and Fiske (1959), no risk of potential biases is induced within the methods, the traits, or a combination of both. The results based on this methodology suggest that our web-based indicators reflect the importance given to innovation factors such as R&D and IP. It is also worth mentioning that the validity of this construct appears to be stronger than all the others attempted in section 4.2.

4.4. Post Hoc Analysis: MTMM Confirmatory Factor Analysis

Some method effects can be induced independently "among the items within or between constructs" (Ortiz de Guinea et al., 2013, p. 839) and "based on the nature of the rater, item, construct, and/or context" (Richardson, Simmering, & Sturman, 2009, p. 766). This could mean that one or more traits do not contribute or contribute negatively to the construct, which could

Table 16. MTMM matrix for R&D_INDEX and IP_INDEX

		Web method		Questionnaire method	
	Traits	WEB_RD	WEB_IP	R&D_INDEX	IP_INDEX
Web method	WEB_RD	–[a]			
	WEB_IP	−0.182	–[a]		
Questionnaire method	R&D_INDEX	0.419**	0.294**	–[a]	
	IP_INDEX	−0.017	0.520**	0.320**	–[a]

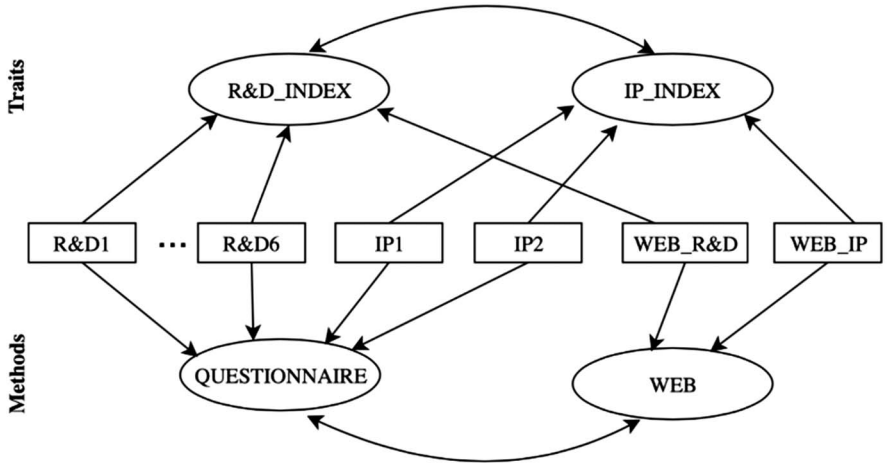
Convergence validity: **Yes**, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**}$ are really high and significant;

Discriminant validity: **Yes**, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**} >> |r_{RD1,IP2}| = 0.017$ and $|r_{IP1,RD2}| = 0.294^{**}$;

Construct validity without either a combination of trait and method effects or a mono-method bias: **Yes**, $r_{RD1,RD2} = 0.419^{**}$ and $r_{IP1,IP2} = 0.520^{**} >> |r_{IP1,RD1}| = 0.182$ and $|r_{RD2,IP2}| = 0.320^{**}$.

Notes: [a]All traits are measured by single items, no reliability statistic can be calculated.

* $p < 0.05$; ** $p < 0.01$



RD: Questionnaire-based survey items measuring R&D; IP: Questionnaire-based survey items measuring IP; WEB_R&D: Web mining item measuring R&D; WEB_IP: Web mining items measuring IP;

Figure 2. MTMM CFA with formative indexes.

ultimately affect the correlations found in the MTMM matrix performed in the previous step. Thus, if these method effects are significant, a congeneric bias is present. One way to verify the effect of each item individually is to extend the MTMM matrix with a CFA, which allows the estimation of latent factors within the MTMM matrix (Maas, Lensvelt-Mulders, & Hox, 2009). Using two PLS regressions allows traits and method factors to load on the items of the estimated constructs. Convergent validity is obtained when the trait loadings are higher than the method loadings. A possible method effect might thus be induced if a method load is higher than the related trait. The discriminant validity is obtained when correlations between the traits are low or moderate. The model tested for this analysis is presented in Figure 2.

MTMM CFA results are presented in Table 17. All the traits load on every item with strong values (all above 0.8). Most of the value loadings for the methods are strong, but generally lower

Table 17. MTMM CFA results

	Items	Trait loading	Method loading
Questionnaire-based survey items measuring R&D	R&D1 (Z_propEmp_R&D)	0.825	0.793
	R&D2 (Z_nb_R&D_proj)	0.921	0.984
	R&D3 (Z_time_R&D)	0.848	0.910
	R&D4 (Z_Ext_R&D_source)	1.000	0.984
	R&D5 (Z_Contract_Out_R&D)	0.983	0.469
	R&D6 (Z_Contract_In_R&D)	0.914	0.905
Questionnaire-based survey items measuring IP	IP1 (Z_nbIP)	0.900	0.993
	IP2 (Z_nbPatent)	1.000	0.978
Web mining item measuring R&D	WEB_R&D (Z_WEB_R&D)	0.996	0.915
Web mining item measuring IP	IP_WEB (Z_WEB_IP)	0.947	0.923

Table 18. MTMM CFA correlations

	R&D_INDEX	IP_INDEX
R&D_INDEX	1.000	
IP_INDEX	0.273*	1.000

* $p < 0.05$.

** $p < 0.01$.

than the trait loading, indicating convergent validity and no congeneric bias. Possible method effects are observed with three items (R&D2, R&D3, and IP1) which load higher with the method than with the trait. It is possible that our use of formative indices may have had an influence on the measure, thus creating a method bias flagged by the CFA results. However, 7/10 (70%) of our results respect the condition, which is considered a high enough proportion to suggest an overall absence of method bias in the construct (Ortiz de Guinea et al., 2013).

Finally, comparing both traits (as set out in Table 18) shows a moderate correlation between the two formative indices, which indicates good discriminant validity.

The results from this post hoc analysis confirm the results obtained using the MTMM matrix alone. Our web-based indicators seem to effectively reflect the importance that technological firms grant to R&D and IP.

5. DISCUSSION AND IMPLICATIONS FOR RESEARCH

We used web content analysis to build innovation indicators from the web text content of four factors that are consequential for the success of high-technology innovation, R&D, IP, collaboration, and external financing. We then validated the true nature of these indicators, by determining whether they were valid substitutes for specific indicators that would otherwise have required a questionnaire-based survey to be obtained. To better understand the nature of the data extracted, two propositions were extensively tested.

Our first proposition stipulated that our web-based indicators could effectively measure specific actions of a company to perform activities related to a given factor. Although significant results were obtained for IP factors and some indicators of R&D, there were no significant correlations with either collaboration or external financing factors.

For the specific case of R&D, we observed that the web-based indicator seems to reflect the promotion needs of various firms in terms of R&D. Our R&D web-based indicator is most highly correlated with the survey-based indicator related to firms more likely to provide R&D services to third parties. Thus, companies use their websites to promote their R&D service offerings. Furthermore, our web-based R&D indicator is highly correlated with the questionnaire-based indicator associated with firms that have a high percentage of employees allocated to R&D tasks. This can be explained by the willingness of firms to attract new R&D talent via their websites. Finally, our R&D web-based indicator correlated significantly with the questionnaire-based indicator tied to whether firms are more likely to contract R&D services from external providers. It is possible that the more important a firm considers R&D, the more likely it is to use both internal and external resources to achieve its R&D-related objectives, which can be described as open innovation practices (Chesbrough, 2003). However, we did not find any correlation with an important R&D indicator, the number of R&D projects and our R&D web-based indicator, which at first may seem counterintuitive, but reflects a limitation of the indicator by not properly accounting for the size dimension of R&D activities.

In contrast, our IP web-based indicator correlates with both the use of IP mechanisms and patent-related activities. Prior to the protection of an intellectual asset, that is, up to the R&D project level, an idea or future invention remains secret. Once an invention is revealed to the world via the disclosure that is generally associated with formal IP protection mechanisms, these IP protection mechanisms act as a signal for the firm, showing how innovative it is or what it wants the world to see. Given that the firms' websites are used to publicly disclosing such information, it thus appears to be a promising source of data for IP protection-related data.

However promising, these preliminary results required more rigorous testing, beyond the simple correlation test between our indicators. The goal was to identify whether specific variables stand out from the others in terms of validity. To do so, we used MTMM matrices to verify the validity of all the promising constructs to assess whether our questionnaire-based indicators can effectively discriminate each other when compared to the web-based indicators. Of the six MTMM matrices produced, four passed all the required validity tests, and two failed.

The fact that we obtained four valid constructs with four different combinations of R&D-IP pairs of different variables demonstrates that we cannot isolate specific actions from the questionnaire-based survey from web-based indicators at this point. Accordingly, at this stage we cannot use the web-based indicator built as a direct substitute for action-related indicators, as good convergent and discriminant validity were obtained for three different questionnaire-based indicators with the R&D web-based indicator and two different questionnaire-based indicators with the IP web-based indicator.

The keywords mentioned on websites may be used by companies that contract R&D services offered by third parties, by companies that provide R&D services to other companies, or by companies that have a higher proportion of employees allocated to R&D, or any combination of the three. We cannot separate one action from another.

The same applies to IP protection. Because acceptable results have been found for the number of patents that a firm owns and for the total number of mechanisms used to protect IP, it is impossible to know precisely what is being measured by our web-based indicator. It is worth mentioning that neither factor significantly correlates with the other ($r = 0.144$; $p > 0.1$), which suggests that both variables represent different traits. Therefore, it is not possible at this point to isolate precise IP-related actions.

The second proposition suggests that our web-based indicators could measure the degree of importance that a company grants to a given factor. To do so, we combined all the elements related to a given factor to create two formative indices from the questionnaire-based indicators: R&D_INDEX and IP_INDEX. We repeated the MTMM analysis using these two indicators and the two web-based indicators and observed the convergent and discriminant validity of our construct. A post hoc analysis with a CFA confirmed our results. Therefore, our proposition 2 is accepted. In other words, we can use our web-based indicators as proxies for the relative importance that a company grants to a given factor (such as R&D and IP). It is possible to conclude that the more R&D-related keywords a firm uses, the more important it considers R&D-related activities to be, irrespective of the nature of the activities. The same also applies to the protection of IP. We know that the more IP protection-related terms are used, the more actions are taken in this direction. However, this methodology does not allow us to precisely determine the nature of the actions a firm takes, which echoes previous findings by (Gök et al., 2015). We can only guess the nature of these actions, as the indicator gives positive answers in both cases because the use of keywords alone ignores the context in which these words are used.

In a nutshell, our methodology can be used as a valid approach to provide data for future innovation and technology management studies for the relative importance given to factors such as R&D and IP, and to test the validity of the measures thus created. In most questionnaire-based surveys, this information is gathered using 1 to 7 Likert scale questions. If the goal of a study is to determine the degree of importance of core factors such as R&D or IP for a firm, the use of web-based indicators is reasonable. However, if the goal is to gather more precise information, such as the specific actions taken by a firm, these web-based indicators lack the necessary context to behave as expected. Although they may provide complementary information, they cannot be used as direct proxies.

The fact that we did not obtain significant correlations with collaboration and external funding suggests erring on the side of caution before using this method on a larger scale. The reasons for the absence of convergent validity of the web-based indicators towards those generated from the questionnaire-based survey can be attributed to the small sample size, the inherent characteristics of the subjects from our sample, the questions used to capture that information within the questionnaire-based survey, and the keywords used to capture the information from the websites. Thus, other tests with different combinations of methods, keywords, and traits need to be explored to determine whether it is possible to obtain valid and relevant information from companies' websites.

6. LIMITATIONS AND FUTURE RESEARCH

Given more data, our research would obviously be more robust, especially in terms of verifying the concept of collaboration and external financing normally addressed by classical methods, which can be appropriately measured using websites. For instance, we were unable to crawl data from all the companies from our survey due to technical limitations, which meant that only 79 out of 89 companies were used in this paper.

Websites are updated from time to time and the information provided changes accordingly, depending on what companies want to make public. It is thus important to note that a single web mining crawl might be insufficient to capture all relevant information, given that results are subject to change as websites are updated. Thus, a longitudinal study would be required to more clearly understand how time can influence the traits to be measured; that is, the content available on a website and the nature of the website itself. An analysis of the evolution of web-based indicators over time would provide further validity to our methodology.

The inability to measure specific actions constitutes a limitation in our web mining methodology. The main problem lies in the lack of context tied to the use of keywords alone, possibly leading to multiple false positives. Machine learning and deep learning techniques, such as recurrent neural network, natural language processing, or bag-of-words models, are promising avenues to explore the necessary context surrounding specific concepts to improve the level of precision of web-based indicators.

Furthermore, we began with theoretical factors for the conceptual framework, then identified the keywords related to these factors, and finally mined the website for these specific keywords. An interesting alternative would be to do this the other way around; in other words, to start with the website content and identify the factors that can be naturally found via unsupervised machine learning algorithms. The term frequency inverse document frequency technique (TF-IDF) could be used to provide insight into the importance of keywords relative to the rest of a document. N-gram low-frequency words clustering could also be tested to better isolate specific areas of interest (Price & Thelwall, 2005).

Company websites are purposely structured in a cooperative and agreeable manner for anyone seeking information about products, services, activities, and so on. The self-reporting bias induced by this methodology is inevitable. However, it is important to note that questionnaire-based surveys and most national official public directories are all also subject to self-reporting biases. Fortunately, the bias induced by the web mining technique is as much a quality as it is a flaw, in that it provides insight on how a company wishes to be perceived. In fact, companies post what they care about, what is important to them, and who they are as an organization on their websites. This qualitative information represents the essence of the company. Future research is needed to determine whether such information could, for instance, be used as a proxy to understand a company's culture.

Other qualitative data analyses of the websites' content could be used to reduce the risk of false positives and to gather more accurate data. The use of indicators based on websites' text data will open the door to experimenting with other possible indicators derived from the same websites, such as the use of colors, pictures, and illustrations; audio and video content; choices of web design styles; use of modern web technologies; frequency of updates; possible interactions with visitors through all of the possible calls to action; and many other sources of data we may not have thought of as yet.

Finding a way to leverage the information from high-tech companies' websites will enable innovation management researchers to access to a whole new source of data that is free to use, accessible at all times, and in large quantities, which will ultimately facilitate studies in this field. Finally, the innovation research community is invited to build on this common ground to create a new systematic way to validate new constructs. If such new innovation indicators from web-based sources can be validated and clearly understood for what they truly measure, the burden on firms that are increasingly asked to answer questionnaires (from researchers, industry associations, or the government), would be considerably reduced. This paper shows that this is a promising avenue.

AUTHOR CONTRIBUTIONS

Mikaël Héroux-Vaillancourt: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Catherine Beaudry: Conceptualization, Methodology, Project administration, Resources, Funding acquisition, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. Constant Rietsch: Resources, Software, Visualization, Writing—original draft.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This research project was supported by Social Sciences and Humanities Research Council grants (number #435-2013-1220, #895-2018-1006) and the Canada Research Chair program.

DATA AVAILABILITY

Data cannot be made available publicly due to the confidentiality contract with the subjects of this study.

REFERENCES

Adams, R., Bessant, J., & Phelps, R. (2006). Innovation management measurement: A review. *International Journal of Management Reviews*, *8*(1), 21–47. **DOI:** https://doi.org/10.1111/j.1468-2370.2006.00119.x

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to "webometrics." *Journal of Documentation*, *53*(4), 404–426. **DOI:** https://doi.org/10.1108/EUM0000000007205

Archibugi, D. (1992). Patenting as an indicator of technological innovation: A review. *Science and Public Policy*, *19*(6), 357–368. **DOI:** https://doi.org/10.1093/spp/19.6.357

Archibugi, D., & Sirilli, G. (2001). The direct measurement of technological innovation in business. In Innovation and enterprise creation: Statistics and indicators. *Proceedings of the Conference Held at Sophia Antipolis*. European Commission (Eurostat), ed. Luxembourg: European Commission.

Armellini, F., Beaudry, C., & Kaminski, P. C. (2017). Open within a box: An analysis of open innovation patterns within Canadian aerospace companies. *Sinergie Italian Journal of Management*, *34*(101), 15–36. **DOI:** https://doi.org/10.7433/s101.2016.02

Armellini, F., Kaminski, P. C., & Beaudry, C. (2014). The open innovation journey in emerging economies: An analysis of the Brazilian aerospace industry. *Journal of Aerospace Technology and Management*, *6*(4), 462–474. **DOI:** https://doi.org/10.5028/jatm.v6i4.390

Arora, A. (1995). Licensing tacit knowledge: Intellectual property rights and the market for know-how. *Economics of Innovation and New Technology*, *4*(1), 41–60. **DOI:** https://doi.org/10.1080/10438599500000013

Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics*, *95*(3), 1189–1207. **DOI:** https://doi.org/10.1007/s11192-013-0950-7

Arvanitis, S. (2012). How do different motives for R&D cooperation affect firm performance?—An analysis based on Swiss micro data. *Journal of Evolutionary Economics*, *22*(5), 981–1007. **DOI:** https://doi.org/10.1007/s00191-012-0273-5

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, *36*(3), 421–458. **DOI:** https://doi.org/10.2307/2393203

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General*, *147*(8), 1264–1272. **DOI:** https://doi.org/10.1037/xge0000383, **PMID:** 30070579

Baysinger, B., & Hoskisson, R. E. (1989). Diversification strategy and R&D intensity in multiproduct firms. *Academy of Management Journal*, *32*(2), 310–332. **DOI:** https://doi.org/10.2307/256364

Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, *26*(5), 644–664. **DOI:** https://doi.org/10.1016/j.technovation.2005.06.016

Belderbos, R., Carree, M., & Lokshin, B. (2004). Cooperative R&D and firm performance. *Research Policy*, *33*(10), 1477–1492. **DOI:** https://doi.org/10.1016/j.respol.2004.07.003

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216–1227. **DOI:** https://doi.org/10.1002/asi.20077

Bozdogan, K., Deyst, J., Hoult, D., & Lucas, M. (1998). Architectural innovation in product development through early supplier integration. *R&D Management*, *28*(3), 163–173. **DOI:** https://doi.org/10.1111/1467-9310.00093

Brown, J. R., Fazzari, S. M., & Petersen, B. C. (2009). Financing innovation and growth: cash flow, external equity, and the 1990s R&D boom. *Journal of Finance*, *64*(1), 151–185. **DOI:** https://doi.org/10.1111/j.1540-6261.2008.01431.x

Campbell, C. M., Michel, J. O., Patel, S., & Gelashvili, M. (2019). College teaching from multiple angles: A multi-trait multi-method analysis of college courses. *Research in Higher Education*, *60*(5), 711–735. **DOI:** https://doi.org/10.1007/s11162-018-9529-8

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. **DOI:** https://doi.org/10.1037/h0046016, **PMID:** 13634291

Carboni, O. A. (2013). Spatial and industry proximity in collaborative research: Evidence from Italian manufacturing firms. *Journal of Technology Transfer*, *38*(6), 896–910. **DOI:** https://doi.org/10.1007/s10961-012-9279-2

Cebon, P., Newton, P., & Noble, P. (1999). Innovation in firms: Towards a framework for indicator development. *Melbourne Business School, Working Paper*, *99-9*.

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Quarterly*, *33*(4), 689–707. **DOI:** https://doi.org/10.2307/20650323

Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Brighton, MA: Harvard Business School Press.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*(s1), 2–9. **DOI:** https://doi.org/10.1111/j.1475-4932.2012.00809.x

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64–73. **DOI:** https://doi.org/10.2307/3150876

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Cohen, W. M., & Levin, R. C. (1989). Empirical studies of innovation and market structure. In *Handbook of Industrial Organization* (Vol. 2, pp. 1059–1107). Amsterdam: Elsevier. **DOI:** https://doi.org/10.1016/S1573-448X(89)02006-6

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, *35*(1), 128–152. **DOI:** https://doi.org/10.2307/2393553

Coombs, R., Narandren, P., & Richards, A. (1996). A literature-based innovation output indicator. *Research Policy*, *25*(3), 403–413. **DOI:** https://doi.org/10.1016/0048-7333(95)00842-X

Crawley, T. (2007). Commercialization of nanotechnology–Key challenges. *Report on the Workshop Organised by Nanoforum*, Helsinki.

Deeds, D. L. (2001). The role of R&D intensity, technical development and absorptive capacity in creating entrepreneurial wealth in high technology start-ups. *Journal of Engineering and Technology Management*, *18*(1), 29–47. **DOI:** https://doi.org/10.1016/S0923-4748(00)00032-1

Diamantopoulos, A. (1999). Viewpoint – Export performance measurement: Reflective versus formative indicators. *International Marketing Review*, *16*(6), 444–457. **DOI:** https://doi.org/10.1108/02651339910300422

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of*

Management, *17*(4), 263–282. **DOI:** https://doi.org/10.1111/j.1467-8551.2006.00500.x
Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277. **DOI:** https://doi.org/10.1509/jmkr.38.2.269.18845
Dosi, G. (1988). Sources, procedures, and microeconomic effects of innovation. *Journal of Economic Literature*, *26*(3), 1120–1171.
Esposito, E. (2004). Strategic alliances and internationalisation in the aircraft manufacturing industry. *Technological Forecasting and Social Change*, *71*(5), 443–468. **DOI:** https://doi.org/10.1016/S0040-1625(03)00002-7
Feldman, M. P., & Florida, R. (1994). The geographic sources of innovation: Technological infrastructure and product innovation in the United States. *Annals of the Association of American Geographers*, *84*(2), 210–229. **DOI:** https://doi.org/10.1111/j.1467-8306.1994.tb01735.x
Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, *112*(3), 393. **DOI:** https://doi.org/10.1037/0033-2909.112.3.393
Flor, M. L., & Oltra, M. J. (2004). Identification of innovating firms through technological innovation indicators: An application to the Spanish ceramic tile industry. *Research Policy*, *33*(2), 323–336. **DOI:** https://doi.org/10.1016/j.respol.2003.09.009
Fosfuri, A. (2006). The licensing dilemma: Understanding the determinants of the rate of technology licensing. *Strategic Management Journal*, *27*(12), 1141–1158. **DOI:** https://doi.org/10.1002/smj.562
Frear, C. R., & Metcalf, L. E. (1995). Strategic alliances and technology networks: A study of a cast-products supplier in the aircraft industry. *Industrial Marketing Management*, *24*(5), 379–390. **DOI:** https://doi.org/10.1016/0019-8501(95)00029-A
Geroski, P., Machin, S., & Van Reenen, J. (1993). The profitability of innovating firms. *RAND Journal of Economics*, *24*(2), 198–211. **DOI:** https://doi.org/10.2307/2555757
Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. **DOI:** https://doi.org/10.1007/s11192-014-1434-0, **PMID:** 26696691, **PMCID:** PMC4677352
Greve, H. R. (2003). A behavioral theory of R&D expenditures and innovations: evidence from shipbuilding. *Academy of Management Journal*, *46*(6), 685–702. **DOI:** https://doi.org/10.2307/30040661
Griliches, Z. (1990). *Patent Statistics as Economic Indicators: A Survey* (Working Paper No. 3301). National Bureau of Economic Research. **DOI:** https://doi.org/10.3386/w3301
Griliches, Z. (1994). Productivity, R&D and the data constraint. Presidential address, American Economic Association, Boston, January 4, 1994. *American Economic Review*, *84*(1), 115–119.
Griliches, Z. (1998). *R&D and productivity*. Chicago, IL: University of Chicago Press. https://ideas.repec.org/b/ucp/bknber/9780226308869.html, **DOI:** https://doi.org/10.7208/chicago/9780226308906.001.0001
Gulek, C. (1999). *Using multiple means of inquiry to gain insight into classrooms: A multi-trait multi-method approach*. https://eric.ed.gov/?id=ED431016
Guo, B., Aveyard, P., Fielding, A., & Sutton, S. (2008). Testing the convergent and discriminant validity of the decisional balance scale of the transtheoretical model using the multi-trait multi-method approach. *Psychology of Addictive Behaviors*, *22*(2), 288–294. **DOI:** https://doi.org/10.1037/0893-164X.22.2.288, **PMID:** 18540726
Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy*, *32*(8), 1365–1379. **DOI:** https://doi.org/10.1016/S0048-7333(02)00137-3
Hagedoorn, J., Link, A. N., & Vonortas, N. S. (2000). Research partnerships. *Research Policy*, *29*(4), 567–586. **DOI:** https://doi.org/10.1016/S0048-7333(99)00090-6
Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5). Upper Saddle River, NJ: Prentice Hall.
Hall, B. H. (1990). The impact of corporate restructuring on industrial research and development. *Brookings*, January 1. https://www.brookings.edu/bpea-articles/the-impact-of-corporate-restructuring-on-industrial-research-and-development/, **DOI:** https://doi.org/10.2307/2534781
Harhoff, D., & Körting, T. (1998). Lending relationships in Germany—Empirical evidence from survey data. *Journal of Banking & Finance*, *22*(10), 1317–1353. **DOI:** https://doi.org/10.1016/S0378-4266(98)00061-2
Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). *Econometric Models for Count Data with an Application to the Patents-R&D Relationship* (Working Paper No. 17). National Bureau of Economic Research. **DOI:** https://doi.org/10.3386/t0017
Haziza, D., & Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, *75*(1), 25–43. **DOI:** https://doi.org/10.1111/j.1751-5823.2006.00002.x
Herrouz, A., Khentout, C., & Djoudi, M. (2013). Overview of web content mining tools. *ArXiv:1307.1024 [Cs]*. http://arxiv.org/abs/1307.1024
Hitt, M. A., Hoskisson, R. E., & Kim, H. (1997). International diversification: Effects on innovation and firm performance in product-diversified firms. *Academy of Management Journal*, *40*(4), 767–798. **DOI:** https://doi.org/10.2307/256948
Hwang, D. (2010). *Ranking the nations on nanotech | Solid State Technology*. http://electroiq.com/blog/2010/08/ranking-the-nations/
Hyun Kim, J. (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. **DOI:** https://doi.org/10.1111/j.1083-6101.2011.01564.x
Johnson, W. H. A., & Filippini, R. (2009). Internal vs. external collaboration: What works. *Research-Technology Management*, *52*(3), 15–17. **DOI:** https://doi.org/10.1080/08956308.2009.11657564
Jordan, J., & Lowe, J. (2004). Protecting strategic knowledge: Insights from collaborative agreements in the aerospace sector. *Technology Analysis & Strategic Management*, *16*(2), 241–259. **DOI:** https://doi.org/10.1080/09537320410001682900
Kalil, T. A. (2005). Nanotechnology and the valley of death. *Nanotechnology Law & Business*, *2*, 265.
Katz, J. S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2), 85–95. **DOI:** https://doi.org/10.3152/147154406781775922
Kim, J., Lee, S., & Marschke, G. (2014). Impact of university scientists on innovations in nanotechnology. In S. Ahn, B. Hall, & K. Lee (eds.), *Intellectual Property for Economic Development* (pp. 141–158). Cheltenham: Edward Elgar Publishing. http://www.elgaronline.com/view/9781782548041.00012.xml, **DOI:** https://doi.org/10.4337/9781782548058.00012
Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). The non-trivial choice between innovation indicators. *Economics of Innovation and New Technology*, *11*(2), 109–121. **DOI:** https://doi.org/10.1080/10438590210899

Klette, T. J., Møen, J., & Griliches, Z. (2000). Do subsidies to commercial R&D reduce market failures? *Microeconometric evaluation studies Research Policy*, *29*(4), 471–495. **DOI:** https://doi.org/10.1016/S0048-7333(99)00086-4

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Laursen, K., & Salter, A. (2006). Open for innovation: The role of openness in explaining innovation performance among U.K. manufacturing firms. *Strategic Management Journal*, *27*(2), 131–150. **DOI:** https://doi.org/10.1002/smj.507

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), 1444. **DOI:** https://doi.org/10.1007/s11051-013-1444-5

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review/Revue Internationale de Statistique*, *54*(2), 139–157. **DOI:** https://doi.org/10.2307/1403140

Lugtig, P. (2017). The relative size of measurement error and attrition error in a panel survey. Comparing them with new multi-trait multi-method model. *Survey Research Methods*, *11*(4), 369–382. **DOI:** https://doi.org/10.18148/srm/2017.v11i4.7170

Maas, C. J., Lensvelt-Mulders, G. J., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology*, *5*(3), 72–77. **DOI:** https://doi.org/10.1027/1614-2241.5.3.72

Mazzoleni, R., & Nelson, R. R. (1998). The benefits and costs of strong patent protection: A contribution to the current debate. *Research Policy*, *27*(3), 273–284. **DOI:** https://doi.org/10.1016/S0048-7333(98)00048-1

McNeil, R. D., Lowe, J., Mastroianni, T., Cronin, J., & Ferk, D. (2007). Barriers to nanotechnology commercialization (pp. 1–57). College of Business and Management, The University of Illinois at Springfield. http://www.wimb.fink.rs/docs/Report-BarriersNanotechnologyCommercialization.pdf

Merges, R. P. (1999). Institutions for intellectual property transactions: The case of patent pools. *University of California at Berkeley Working Paper*, 1–74.

Meuleman, M., & De Maeseneire, W. (2012). Do R&D subsidies affect SMEs' access to external financing? *Research Policy*, *41*(3), 580–591. **DOI:** https://doi.org/10.1016/j.respol.2012.01.001

Michie, J. (1998). Introduction. The Internationalisation of the Innovation Process. *International Journal of the Economics of Business*, *5*(3), 261–277. **DOI:** https://doi.org/10.1080/13571519884387

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. New York: Academic Press. **DOI:** https://doi.org/10.1016/B978-0-12-386979-1.00020-7, https://doi.org/10.1016/B978-0-12-386979-1.00026-8

Minguillo, D., & Thelwall, M. (2012). Mapping the network structure of science parks: An exploratory study of cross-sectoral interactions reflected on the web. *Aslib Proceedings*, *64*(4), 332–357. **DOI:** https://doi.org/10.1108/00012531211244716

National Nanotechnology Coordination Office. (2017). *Supplement to the President's 2018 Budget* (p. 86).

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45–65. **DOI:** https://doi.org/10.1016/S0169-7439(96)00007-X

OECD & Statistical Office of the European Communities. (2005). *Oslo Manual*. https://www.oecd-ilibrary.org/content/publication/9789264013100-en

OECD & Eurostat. (2019). *Oslo Manual 2018*. https://www.oecd-ilibrary.org/content/publication/9789264304604-en

Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2013). Measure for measure: A two study multi-trait multi-method investigation of construct validity in IS research. *Computers in Human Behavior*, *29*(3), 833–844. **DOI:** https://doi.org/10.1016/j.chb.2012.12.009

Parker, H. (2000). Interfirm collaboration and the new product development process. *Industrial Management & Data Systems*, *100*(6), 255–260. **DOI:** https://doi.org/10.1108/02635570010301179

Parthasarthy, R., & Hammond, J. (2002). Product innovation input and outcome: Moderating effects of the innovation process. *Journal of Engineering and Technology Management*, *19*(1), 75–91. **DOI:** https://doi.org/10.1016/S0923-4748(01)00047-9

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: Possibilities and problems. *Scientometrics*, *7*(1–2), 77–99. **DOI:** https://doi.org/10.1007/BF02020142

Peter, J. P., & Churchill, G. A. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research*, *23*(1), 1–10. **DOI:** https://doi.org/10.2307/3151771

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly*, *31*(4), 623–656. **DOI:** https://doi.org/10.2307/25148814

Price, L., & Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology*, *56*(8), 883–888. **DOI:** https://doi.org/10.1002/asi.20177

Ramdani, A. (2014). *Revue systématique de la littérature sur les mesures de la collaboration inter-organisationnelle dans un contexte d'innovation* [Masters, École Polytechnique de Montréal]. https://publications.polymtl.ca/1624/

Reinig, B. A., Briggs, R. O., & Nunamaker, J. F. (2007). On the measurement of ideation quality. *Journal of Management Information Systems*, *23*(4), 143–161. **DOI:** https://doi.org/10.2753/MIS0742-1222230407

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods*, *12*(4), 762–800. **DOI:** https://doi.org/10.1177/1094428109332834

Rivette, K. G., & Kline, D. (2000). *Rembrandts in the attic: Unlocking the hidden value of patents*. Boston, MA: Harvard Business School Press.

Roja, A. I., & Nastase, M. (2013). Leveraging organizational capabilities through collaboration and collaborative competitive advantage. *Revista de Management Comparat International*, *14*(3), 359–366.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer. **DOI:** https://doi.org/10.1007/978-1-4612-4378-6

Straub, D., & Burton-Jones, A. (2007). Veni, vidi, vici: Breaking the TAM logjam. *Journal of the Association for Information Systems; Atlanta*, *8*(4), 223–229. **DOI:** https://doi.org/10.17705/1jais.00124

Straub, D., Limayem, M., & Karahanna-Evaristo, E. (1995). Measuring system usage: Implications for IS theory testing. *Management Science*, *41*(8), 1328–1342. **DOI:** https://doi.org/10.1287/mnsc.41.8.1328

Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK West Midlands

automobile industry. *Research Evaluation*, *15*(2), 97–106. **DOI:** https://doi.org/10.3152/147154406781775968
Teece, D. J. (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy*, *15*(6), 285–305. **DOI:** https://doi.org/10.1016/0048-7333(86)90027-2
Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1–116. **DOI:** https://doi.org/10.2200/S00176ED1V01Y200903ICR004
Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418. **DOI:** https://doi.org/10.1002/asi.21462
Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, *4*, 278–283.
Van de Lei, T. E., & Cunningham, S. W. (2006). Use of the internet for future-oriented technology analysis. *2nd International Seville Seminar on Future-Oriented Technology Analysis: Impact of FTA Approaches on Policy and Decision-Making* (pp. 28–29). Seville, Spain.
Vaughan, L. (2004). Exploring website features for business information. *Scientometrics*, *61*(3), 467–477. **DOI:** https://doi.org/10.1023/b:scie.0000045122.93018.2a
Weare, C., & Lin, W.-Y. (2000). Content analysis of the world wide web: Opportunities and challenges. *Social Science Computer Review*, *18*(3), 272–292. **DOI:** https://doi.org/10.1177/089443930001800304
Youtie, J., Hicks, D., Shapira, P., & Horsely, T. (2012). Pathways from discovery to commercialization: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. *Technology Analysis & Strategic Management*, *24*(10), 981–995. **DOI:** https://doi.org/10.1080/09537325.2012.724163

APPENDIX 1: RELEVANT 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?

   - Applied Research;
   - Product Scoping and Business Case Building;
   - Development, Testing and Validation;
   - 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 commercialisation 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

8. Did your firm participate in alliances or collaborative agreements with other organisations to develop or commercialise your MSR product innovation? Y/N
9. 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

10. 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.

	Development of innovation	Commercialisation of innovation
Internal funds of your firm or establishment		
Government subsidies/tax credits/academics grants		
Debt capital (such as bank loans)		
Venture capital (public/private)		
Collaboration agreements		
Programs from organisations such as nanoQuebec (now PRIMA-Quebec), nanoOntario, nanoAlberta, etc.		
Other		

Intellectual property

11. 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

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

	All patents	Nanotechnology-related and advanced materials patents
Patent applications		
Existing patents		
Patents assigned/(sold) to others		

APPENDIX 2: CORRELATIONS AND DESCRIPTIVE STATISTICS

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
WEB_R&D	1	1.00																		
WEB_IP	2	−0.18	1.00																	
WEB_Collab	3	0.35**	−0.30**	1.00																
WEB _ExtFin	4	0.23*	−0.37**	0.42**	1.00															
ln(nb_R&D_proj)	5	0.00	−0.13	−0.00	0.04	1.00														
dInt_R&D_source	6	0.17	0.01	0.08	0.05	0.09	1.00													
Ext_R&D_source	7	0.15	0.20	0.09	0.07	0.19	−0.07	1.00												
Contract_In_R&D	8	0.28**	−0.21*	0.28**	0.15	0.14	−0.06	0.40**	1.00											
Contract_Out_R&D	9	0.31**	−0.07	0.08	0.05	−0.05	0.10	−0.11	0.06	1.00										
ln(time_R&D)	10	0.20*	−0.21*	0.24*	0.17	0.10	−0.09	0.16	0.25*	0.07	1.00									
propEmp_R&D	11	0.31**	−0.46**	0.12	0.39**	−0.07	−0.17	0.01	0.21*	0.17	0.31**	1.00								
dCollab_R&D	12	−0.02	−0.19	0.12	0.14	0.10	−1.00**	0.21	0.01	−0.24	0.12	0.21	1.00							
dCollab_R&D_Reason	13	−0.27	−0.02	0.09	−0.15	0.11	−0.23	0.37*	0.22	−0.18	0.57**	−0.13	0.23	1.00						
nbIP	14	0.01	−0.37**	0.23*	0.38**	0.24*	−0.12	0.17	0.43**	−0.07	0.26*	0.28**	0.20	0.27	1.00					
ln(nbPatent)	15	0.04	−0.40**	0.09	0.01	0.09	0.33*	−0.12	−0.08	−0.13	0.18	0.02	−0.24	0.16	0.14	1.00				
dCollab	16	0.17	−0.25*	0.22*	0.11	0.19	0.06	−0.03	0.22*	0.07	0.14	0.11	.a	.a	0.22*	0.13	1.00			
propExtFin_R&D	17	0.36**	−0.18	0.31**	0.15	0.16	−0.07	0.07	0.23*	0.10	0.18	0.24*	0.19	−0.28	0.14	−0.23	0.36**	1.00		
propExtFin_Comm	18	0.37**	−0.30**	0.28**	0.22*	0.11	0.01	0.03	0.26*	0.24*	0.11	0.16	−0.17	−0.06	0.20*	−0.09	0.29**	0.675**	1.00	
propTotExtFin	19	0.40**	−0.27*	0.33**	0.19	0.15	−0.04	0.07	0.26*	0.17	0.16	0.21*	−0.01	−0.18	0.18	−0.16	0.38**	0.91**	0.92**	1.00
N		78.00	78.00	78.00	78.00	84.00	84.00	88.00	86.00	86.00	80.00	79.00	37.00	31.00	80.00	44.00	76.00	80.00	77.00	77.00
Mean		2.43	0.58	1.90	0.27	10.67	0.95	3.90	3.47	3.77	38.06	33.85	0.03	0.42	3.23	17.70	0.50	44.09	26.31	34.63
Std. Deviation		3.32	1.69	2.82	0.59	15.36	0.21	1.78	1.81	2.14	28.61	30.70	0.16	0.50	1.28	33.27	0.50	33.69	35.02	31.39

[a] Cannot be computed because at least one of the variances is constant.

* $p < 0.05$; ** $p < 0.01$

APPENDIX 3: CORRELATIONS BETWEEN INDIVIDUAL QUESTIONNAIRE-BASED AND WEB-MINING INDICATORS

	I	II	III	IV
Variables	WEB_R&D	WEB_IP	WEB_Collab	WEB_ExtFin
ln(nb_R&D_proj)	0.002			
dInt_R&D_source	0.171			
Ext_R&D_source	0.152			
Contract_In_R&D	0.284**			
Contract_Out_R&D	0.306***			
ln(time_R&D)	0.197*			
propEmp_R&D	0.306***			
dCollab_R&D	−0.017			
dCollab_R&D_Reason	−0.270			
nbIP		−0.368***		
ln(nbPatent)		−0.396**		
dCollab			0.222*	
propExtFin_R&D				0.150
propExtFin_Comm				0.222*
propTotExtFin				0.192

Notes: Group I, II, III performed with Pearson correlations and IV performed with Spearman correlations. Two-tailed test.

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

APPENDIX 4: PLS REGRESSIONS

PLS Model B Basic Linear Stable 3 for INDEX_R&D

Items	Weights	SE	*p*	VIF	WLS	ES
Z_propEmp_R&D	0.323	0.101	0.001	1.229	1	0.166
Z_nb_R&D_proj	0.323	0.101	0.001	1.087	1	0.129
Z_time_R&D	0.323	0.101	0.001	1.255	1	0.206
Z_Ext_R&D_source	0.323	0.101	0.001	1.346	1	0.175
Z_Contract_Out_R&D	0.323	0.101	0.001	1.071	1	0.122
Z_Contract_In_R&D	0.323	0.101	0.001	1.352	1	0.201

PLS Model B Basic Warp3 Basic Stable 3 for INDEX_R&D

Items	Weights	SE	*p*	VIF	WLS	ES
Z_propEmp_R&D	0.323	0.101	0.001	1.229	1	0.166
Z_nb_R&D_proj	0.323	0.101	0.001	1.087	1	0.129
Z_time_R&D	0.323	0.101	0.001	1.255	1	0.206
Z_Ext_R&D_source	0.323	0.101	0.001	1.346	1	0.175
Z_Contract_Out_R&D	0.323	0.101	0.001	1.071	1	0.122
Z_Contract_In_R&D	0.323	0.101	0.001	1.352	1	0.201

PLS Model B Basic Warp3 Basic Stable 3 for INDEX_IP

Items	Weights	SE	*p*	VIF	WLS	ES
Z_nbIP	0.636	0.093	<0.001	1.059	1	0.5
Z_nbPatent	0.636	0.093	<0.001	1.059	1	0.5

PLS Model B Basic Linear Basic Stable 3 for INDEX_IP

Items	Weights	SE	*p*	VIF	WLS	ES
Z_nbIP	0.636	0.093	<0.001	1.059	1	0.5
Z_nbPatent	0.636	0.093	<0.001	1.059	1	0.5