funcas

SOCIAL INNOVATION AMONG SPANISH INVENTORS: AN EXPLORATORY ANALYSIS

Mariano Mastrogiorgio & Martina Pasquini

Febrero 2023

Social innovation among Spanish inventors: An exploratory analysis

Mariano Mastrogiorgio & Martina Pasquini[1]

Febrero 2023

Despite the growing interest in social innovation, the literature on this topic is limited both at the theoretical and empirical level. Accordingly, a controversial issue concerns the lack of a shared definition of social innovation and, consequently, the existence of a validated method to classify and measure it. In this paper, we focus on one type of intellectual property rights (IPRs), namely, patents, which well-reflect firms' intangible assets. We propose an exploratory approach based on natural language processing to identify patents with a social-oriented content that can be assimilated to social innovation. We employ a random sample of about 3800 U.S. patents of Spanish inventors to test our methodology and provide novel evidence. Our main findings show that first, social innovation in patents is multi-faced but mostly concentrated in green technologies; and second, that patents with a high level of social content are the ones with a significantly higher number of forward citations—that is, more radical and likely with more value.

One of the most impactful news during the 2019 Business Roundtable was the fact that the leaders of some American biggest companies changed their corporations 'purpose statement', which shifted from addressing exclusively shareholders' profits to considering 'all stakeholders'—that is, employees, customers, and different social actors (Wall Street Journal, 2019). Similarly, the European Union proposed that "for-profit parts of the private sector would need to be further encouraged to use the potential of social investment" (European Union, 2013: pg. 5). One way to reach these broad goals is through *social innovation*, which has received a rapidly growing interest during the last years, both from an academic and policy viewpoint (Cajaiba-Santana, 2014; Dawson and Daniel, 2010; Lee, 2014; Nicholls and Murdock, 2011; Pol and Ville, 2009; van der Have and Rubalcaba, 2016; Pel *et al.*, 2020). The concept is still fraught with conceptual ambiguity, and a plurality of definitions still exists; yet there is wide agreement that "*social innovation encompasses change in social relationships*" that "*serve a shared human need/goal or solve a socially relevant problem*" (van der Have and Rubalcaba, 2016: pg. 1930).

Despite this growing interest, the theoretical and empirical literature on social innovation is limited and scattered. To the best of our knowledge, in this debate, an issue that deserves attention is whether the stimuli and incentives that a firm receives to invest into 'social innovation' can be reflected in a second stage into its intangible value like, for instance, in the value of its *patents,* which represent one formally protected intellectual property rights (IPRs). This focus acquires relevance if we consider that social innovations might have peculiar features, different from standard forms of innovation and, according to some scholars, they can fit less with standard forms of legal protection (Lee, 2014). In this line, a significant limiting factor in order to address such issue is the lack of a validated method to identify and classify social innovation in patents.

To fill this gap, we propose a patent-level approach based on 'natural language

[1] IE University—Department of Strategy
Paseo de la Castellana 259, 28046 Madrid

processing' (Arts *et al.*, 2021; Kaplan and Vakili, 2015; Mitkov, 2004). First, from the Patents View database we gathered data on the patenting activity of *Spanish inventors* in the U.S. patent system—considered as one of the world's largest and valuable (Hall *et al.*, 2001; Kuhn and Thompson, 2019; Kuhn *et al.*, 2020). Second, from Google Patents, we text-mined and then classified the patent text documents, using a well-defined social dictionary developed by Corporate Knights, a research company established in 2002 aiming to promote a 'clean capitalism' that meets social and environmental objectives. After identifying a random sample of the patents, we ran a series of analysis to understand the characteristics of these social-oriented patents. Our findings suggest that first, social innovation has different facets and belongs to different clusters (in line with the literature: see van der Have and Rubalcaba, 2016), and second, that social innovations have a technological impact, as it is associated to a higher number of patent citations. Although our analysis still has an explorative nature (rather than causal), it offers some important evidence that aims to stimulate future theoretical investigations and large-scale implementations of the proposed approach.

Social innovation

Social innovation has received a rapidly growing interest during the last years. Given that a systematic review goes beyond the scope of this paper, we refer to other works on the topic (Cajaiba-Santana, 2014; Dawson and Daniel, 2010; Lee, 2014; Nicholls and Murdock, 2011; Pel *et al.*, 2020; Pol and Ville, 2009; van der Have and Rubalcaba, 2016). A recent and comprehensive summary of the literature can be found in van der Have and Rubalcaba (2016), who traced the history, themes and trends of social innovation based on a bibliographic analysis of scholarly articles in innovation journals.

A key finding of the analysis is the lack of integration and a certain degree of conceptual ambiguity in the social innovation field, due to the complex history of the concept (Defourny and Develtere, 1999; Mulgan *et al.*, 2007), which alternates between sociological and economic interpretations. While sociologically oriented accounts (Cajaiba-Santana, 2014) tend to emphasize the 'processes' of creating social innovation through "change in social relationships", economic oriented accounts (Pol and Ville, 2009) would rather emphasize the 'outcomes' of social innovation, like new technologies, products, or services that "serve a shared human need/goal or solve a socially relevant problem" (van der Have and Rubalcaba, 2016: pg. 1930).

Interestingly, this diversity of interpretations reflects into the existence of distinct, yet interrelated, clusters of social innovation (van der Have and Rubalcaba, 2016): i) one first cluster that refers 'social and societal challenges', which is outcome-oriented and covers themes such as climate sustainability, natural environment and health provisions (Chataway *et al.*, 2010; Markard *et al.*, 2012); ii) a second 'local development' cluster, more process-oriented that covers themes such as the empowerment of citizens and local communities, the role of institutions and governance and different forms of social

cohesion (Moulaert *et al.*, 2005; Nelson and Sampat, 2001); then, a iii) third 'community psychology' and iv) a final 'creativity research' clusters, which are less prominent, both process-oriented, related to the behavioural and creative underpinning of social change (Hazel and Onaga, 2003; Mumford, 2002).

Overall, the literature depicts a complex portray of social innovation, whose peculiar features require a holistic interpretative framework. This point has been raised also in some streams of the patent literature, which we review briefly below.

Social innovation in the patent literature

According to standard interpretations, *innovation* should qualify itself as a public good: therefore, non-rivalry and non-excludability make it difficult to exclude others from replication and use. This partial appropriability of returns, which covers the costs of knowledge production, would create low incentives to innovate and, accordingly, it would lead to market failure in the form of under-production of innovation (Arrow, 1962; Dosi *et al.*, 2006). In this context, patents aim to re-establish the appropriability of returns and, by implication, the incentives to innovate by granting a temporal monopoly to novel, non-obvious, disclosed knowledge, (Gallini and Scotchmer, 2002) and by providing the public a 'notice' of intellectual property boundaries (Bessen and Meurer, 2008). All in all, this suggests a "narrow, particularized conception" according to which innovation is "individualistic, discrete, novel, and objectively reproducible" (Lee, 2014: pg. 4).

On a different approach, some scholars pinpoint that social innovation challenges this standard interpretation of innovation in two ways (Henry and Stiglitz, 2010; Lee, 2014). On the one hand, social innovators are driven by the quest for social impact, which reduces the problem of appropriability of returns, and thus the under-production due to diminished incentives (Desmarchelier *et al.*, 2020). On the other hand, social innovation exhibits a range of features that distance it from the narrow conception (Lee, 2014), contrary to the idea of innovation as individualistic, discrete, novel, and objectively reproducible. In this respect, social innovation often comes from distributed, collaborative social communities, rather than being individualistic (Montgomery *et al.*, 2012); it is somewhat amorphous in time and boundaries, rather than being discrete (Lemley, 2013); its value often comes from applying what has already been done successfully, rather than being novel *stricto sensu* (Hargadon, 2003); its reproduction, extension and diffusion requires dense community networks, thus diverging from the "conception of readily and objectively reproducible technology", according to which a legally disclosed patent enables a "person of ordinary skill in the art" to build on it, and extend it, in novel and non-obvious ways (Lee, 2014: pg. 37; De Jong *et al.*, 2015; Jeppesen, 2021).

This raises the question of whether social innovation could reflect into the intangibles formally protected by intellectual property rights—like patents, for instance. Debating if patents benefit or harm society, or if regimes of exclusivity are appropriate or not for social

innovation, clearly goes well beyond the scope of this paper, and it has been covered elsewhere (Bessen and Meurer, 2008; Boldrin and Levine, 2008; Cimoli *et al.*, 2014; Moser, 2013; Ziedonis, 2008). We rather aim to present some exploratory evidence, with the hope of inspiring theory building and future empirical studies on social innovation, proxied in patents with social-oriented content.

An exploratory analysis

Patents represent a key data source in the innovation literature (Hall *et al.*, 2001). For the objective of this study, we gathered patent data from Patents View[2], a database focusing on intellectual property (IP) data that was launched in 2017 in collaboration with the USPTO. Through the advanced query builder, we downloaded all the U.S. granted patents containing at least one Spanish inventor, including the disambiguated name, surname, and city of the inventor. After some further manual cleaning of the data aimed at removing ambiguous cases, we obtained a database containing 20288 patents, granted between 1976 and 2022. **TABLE 1** reports some descriptive statistics, such as the average number of Spanish inventors in U.S. patents (2.198) and the tabulation across count values (50% of U.S. patents contain just one Spanish inventor)[3].

From these 20288 patents, we extracted a random sample of about 5000 patents[4], for which we mined and classified the different elements of patent text, based on the social dictionary of reference; namely, the one elaborated by Corporate Knights. We report details of the text-mining in the next paragraph. Once analysed the patents' text to identify whether they contain social content, we merged these data with other patent sources, such as patents' bibliographic data (obtained from the USPTO Custom Data DVD available for download, available until 2014), forward citations (obtained from Kuhn *et al.*, 2020, observed until 2017 to deal with truncation), and other sources explained after. Our final matched dataset used in the regression models consists of a random sample of 3846 patents, granted between 1976 and 2014 and spanning 353 different technological classes.

The measure

To the best of our knowledge, studies of social innovation employing patent-level data are still limited and case-specific (Berrone *et al.*, 2013; Nameroff *et al.*, 2004). Here we fill this gap by proposing a measure of social innovation in patent documents based on natural language processing, that is, a set of methods in computational linguistics (Mitkov, 2004) adopted in the innovation and patent literature for processing text in large amounts of documents (e.g. Arts *et al.*, 2021; Arts *et al.*, 2018; Balsmeier *et al.*, 2018; Bergeaud *et al.*, 2017; Gerken and Moehrle, 2012; Kaplan and Vakili, 2015; Teodoridis *et al.*, 2020; Von Graevenitz *et al.*, 2021). To achieve our

[2] https://patentsview.org/

[3] The five most active Spanish cities in terms of U.S. patenting are the following: Barcelona, Madrid, Sant Cugat del Vallés, Valencia, Sevilla.

[4] More precisely, the algorithm extracted 4995 patents.

purpose and identify the eventual social content of the patent, in a Matlab environment (Text Analytics Toolbox) we iteratively text-mined patent documents from Google Patents. After parsing the HTML code, we extracted the target text components (abstract, description and claims) and passed them through standard filters in order to remove stop words and short words, reduce words to their root form, erase punctuations, and process the text for part-of-speech tagging. *Social innovation*, our independent variable, is given by *word count*—that is, the frequency of social-innovation-related words in the patent text. The choice of a social dictionary is key: since the quality of text mining depends on the quality of the dictionary, researchers often rely on well-established, available dictionaries (Deng *et al.*, 2018; Morris, 1994). For our purposes, we relied on a rich social dictionary developed by Corporate Knights, whose entries and social meaning are widely shared.

The social dictionary. Corporate Knights is a Canadian company established in 2002 that aims to promote a 'clean capitalism' that meets environmental and social objectives and is well known for its quarterly magazine on sustainable business. Besides the media division, the company owns a research division that produces different types of sustainability rankings of leaders, corporations, stock exchanges and MBAs. The 'Better World MBA Ranking', for instance, ranks 40 MBA programs drawn either from the Financial Times list of the 100 best global MBAs or from the list of educational institutions that signed the 'United Nations Principles for Responsible Management Education', or that simply appeared into the previous-year ranking of Corporate Knights. To produce the ranking, each MBA is evaluated on several sustainability-performance indicators based on publicly available information, like core courses, faculty publications and activities of research institutes, by looking at the integration into these items of a list of sustainability topics, which constitutes a social dictionary made of 386 entries[5].

Fine-tuning the social dictionary. Although the social dictionary has been already validated by Corporate Knights, we processed it further, with the aim of producing a simplified vector of unique words. First, we removed common words (like 'and'), acronyms (like 'VCS') and separators (like '/'), and those few words that did not appear in the multi-dimensional Fast Text Word Embedding vector space (explained in the next section). Second, we analysed each word with Sketch Engine, a text analysis software developed by Lexical Computing Ltd. that is commonly used by linguists, lexicographers, and translators. A key tool of the software is Word Sketches, an application that, given a word input (e.g., 'first'), outputs the words that most frequently co-appear in millions of English-language text corpora ('time', 'one', 'name'). As in 'distributional semantics' (Lappin and Fox,

[5] https://www.corporateknights.com/

2015), the philosophy of this approach is epitomized by J.R. Firth's quote: "you shall know a word by the company it keeps". That is, the aim was that of excluding (from the dictionary) those words with more generic—rather than social-specific—co-occurrences and thus linked to 'semantic' contexts that were too broad for the purposes of our analysis. To further process the dictionary, we hired ten human classifiers on the Amazon's Mechanical Turk, with the aim of excluding other redundant words. That is, for each entry of the dictionary resulting from the previous step, we presented a simple word classification task to each subject: 'does this word refer to social innovation?', where 'social innovation' was defined according to the 17 Sustainable Development Goals (SDGs) of the United Nations, which were shown in the window of each classification task[6]. If the majority of subjects did not classify a word as social, the word was excluded from the dictionary. Through these steps, the initial dictionary has been reduced to a vector of 267 unique words.

Word clustering. Based on a bibliographic analysis of academic articles, van der Have and Rubalcaba (2016) show that the field of social innovation is not homogeneous, as it consists of four macro thematic clusters: environmental sustainability, local development, promotion of behavioural change towards society, creativity in social innovation processes. To capture these different dimensions of social innovation, we clustered accordingly the words of our social dictionary, with the aim of calculating cluster-specific word frequencies in patent text. To cluster words, we mapped each word to its corresponding vector in a multi-dimensional Fast Text Word Embedding space, and then split the words into clusters through a k-means approach guided by the optimization of a Silhouette function to determine the optimal number of clusters (Lengyel and Dukat, 2019). Fast Text Word Embedding is a specific 'word embedding' developed by Facebook AI Research (Bojanowski *et al.*, 2017), which converts each word of the English dictionary (for a total of about one million words) to a vector in a multi-dimensional space of 300 hundred dimensions capturing latent semantic meanings. The vector (expressing the word's position in the semantic space) is learned through a self-supervised approach fed by the co-occurrence of words in millions of text corpora. That is, words that tend to co-occur in text corpora tend to be closer to each other in the multi-dimensional semantic space, thus forming clouds of points that can be split into clusters. The word clusters are reported in **TABLE 2**. As we can see in the table, there are four main clusters, consistently with the findings of van der Have and Rubalcaba (2016). While the first two clusters clearly relate to environmental sustainability, the other two clusters seem to be related to local development and pro-active social behaviours, although their composition is less clear.

[6] https://sdgs.un.org/goals

Key findings

Some descriptive statistics of social innovation are reported in **TABLE 3**. The table reports the aggregated word count (*word count*), the size of the mined text (*doc size*) and the ratio between the two (*ratio total*) in the whole patent and, separately, in the abstract, description and claims, followed by the ratios of cluster-specific word counts (see also Table 2). As shown in Table 3, word counts are significantly skewed (which is quite common in patent data). As shown in the last column of the table, at the right extreme of the word count distribution up to 9.7% of the mined text has a social nature (based on the ratio total), a percentage that changes to 15%, 8.3% and 7.7% in the abstract, description, and claims (based on the *ratio abstract*, *ratio description* and *ratio claims*, respectively). In the second part of the table, the five patents with the highest word count (as expressed by ratio total) are also reported. As we can see, some of them clearly relate to green technology and sustainability, one of the key dimensions of social innovation.

In order to further examine the nature of words counts, we upgraded the algorithm with the aim of identifying the keywords from the social dictionary in the abstract, description and claims of each patent. After converting the word columns into a 'document-feature matrix' (also known as the 'bag of words': see Grimmer *et al.*, 2022), we plotted the word clouds of abstract, description, and claims, reported in **TABLE 4**.

A simple model

Are social innovations impactful from a technological viewpoint, as reflected in a higher number of patent citations?

Below we report some concluding results of a simple model in which we regress (under a negative binomial specification) the number of forward citations received by a patent on the word counts explained in the previous sub-section. Assessing the technological impact of a patent, and how such impact varies in function of social innovation, is a complex undertaking, due to a mixture of monetary and societal dimensions. A well-established proxy of the technological impact of a patent is its number *forward citations*, employed as our dependent variable (Corredoira and Banerjee, 2015). We obtained forward citations from Kuhn *et al.* (2020), and citations to pre-grant publications were also included in the count, following the 1999 American Inventors Protection Act (AIPA) that allowed to consider them as prior art.

Despite their established use, forward citations require corrections for systematic sources of variation across time and technological classes (Lerner and Seru, 2017). Therefore, in all the models, we included *technological class* and *grant year dummies*. Besides the dummies, we added other controls. At the technology level, we controlled for *backward citations*, to assess the novelty (or, inversely, the incremental nature) of the invention, which could drive both social impact and the number of forward citations; moreover, we aimed to remove systematic sources of variation in citations, given that a general increase in

citations could be driven by a few patents citing many other patents (Kuhn *et al.*, 2020). We also controlled for *non-patent citations*, that is, citations to academic articles, because patents that cite science may contain more basic knowledge that may increase their generality (Cassiman *et al.*, 2008), which could in turn drive both social innovation and the number of forward citations. At the inventive-entity level, we controlled for the *number of inventors*. At the assignee level, we introduced a dummy that identifies whether the assignee is a *large entity* (Alcacer *et al.*, 2009). Finally, since higher word counts tend to occur in larger documents, we controlled for *doc size* in all the models.

The concluding results are reported in **TABLE 5**. In the first panel of the table, Model 1 reports the coefficients of word count and doc size, Model 2 adds technological class and grant year dummies, and Model 3 reports the full model. In the second panel of the table, Models 4, 5 and 6 respectively decompose the word count across the abstract, description and claims of a patent. As we can see in all the models, the coefficient of word count is positive and significant at the 1% or 5% level. This shows that, in a random sample of U.S. patents of Spanish inventors, social innovation correlates with downstream technological impact. Needless to say, our analysis is primarily explorative rather than causal, yet we offer some evidence that aims to stimulate future theoretical investigations and large-scale implementations of the proposed approach.

Conclusions

The literature on social innovation is controversial, since at the theoretical level there is not a univocal agreement on what social innovation is and, at the same time, at the empirical level it is not established how to identify and quantify it. The aim of this work has been that of focusing on patents, as one of the main formal firms' intellectual property rights that can represent the intangible value of companies, and of identifying the ones that can proxy social innovation and exploring their characteristics. We employed an exploratory approach based on natural language processing and we tested it on a random sample of about 3800 U.S. patents of Spanish inventors. Our main findings confirm that social innovation is a multi-faced construct which addresses broad social-oriented needs, but patents are typically used to protect 'green technologies'. At technological level, social-oriented patents have a significant impact with respect to regular patents, as shown by their higher number of forward citations.

Although our work still has an explorative nature (rather than a causal one), to the best of our knowledge it is one of the first attempts to measure quantitatively social innovation. On the one hand, this study calls for further investigations and refined techniques also with the use of machine learning (ML) that can help to disentangle the social content of patents and their value for firms (Miric *et al.*, 2023). On the other hand, this study enters into a larger debate on how we can quantify the impact of socially responsible actions and the metrics

that should be used to assess the facets of a firm's ESG (i.e., Environmental-Social-Governance) strategy, which would signal to the market the quality of the social actions (and/or attract investments). In this respect, while some measures are nowadays quite standard, like those related to carbon footprint, wastewater, and paper-recycling, the ones that are related to (broader) social and community-oriented content are more subject to ambiguity and, accordingly, accreditation and rating agencies can set the standards. In this regard, it is not surprising that companies declaring to have higher social innovation tend to have a certification to grant it publicly. Finally, the need of measuring social innovation is key for identifying the characteristics of those social investments that drive real innovation and create value. This has key implications for both researchers and policy makers, who can distinguish it from greenwashing, and thus identify stakeholders that authentically do "walk the talk" and accordingly deserve more public (or private) funds, and legitimacy in front of stakeholders' eyes.

Acknowledgments

For comments and suggestions, we thank Marco Giarratana and numerous participants at the Academy of Management, Strategic Management Society, and B Academics Spain conferences, and at seminars at University of Navarra, IE University, and University of Padova. This paper is based on a larger version containing additional analyses, which is available upon request from the authors (mail to: mmastrogiorgio@faculty.ie.edu).

References

Alcacer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy*, *38(2)*, 415-427.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors* (pp. 609-626). Princeton University Press.

Arts, S., Cassiman, B., & Gomez, J.C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, *39*(1), 62-84.

Arts, S., Hou, J., & Gomez, J.C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*(2), 104144.

Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., ... & Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy*, *27*(3), 535-553.

Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. *PloS One*, *12*(4), e0176310.

Berrone, P., Fosfuri, A., Gelabert, L., & Gomez-Mejia, L.R. (2013). Necessity as the mother of 'green' inventions: Institutional pressures and environmental innovations. *Strategic Management Journal*, *34*(8), 891-909.

Bessen, J., & Meurer, M.J. (2009). *Patent failure*. Princeton University Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Boldrin, M., & Levine, D.K. (2008). *Against intellectual monopoly*. Cambridge University Press.

Cajaiba-Santana, G. (2014). Social innovation: Moving the field forward. A conceptual framework. *Technological Forecasting and Social Change*, *82*, 42-51.

Cassiman, B., Veugelers, R., & Zuniga, P. (2008). In search of performance effects of (in) direct industry science links. *Industrial and Corporate Change*, *17*(4), 611-646.

Chataway, J., Hanlin, R., Mugwagwa, J., & Muraguri, L. (2010). Global health social technologies: Reflections on evolving theories and landscapes. *Research Policy*, *39*(10), 1277-1288.

Cimoli, M., Dosi, G., Maskus, K.E., Okediji, R.L., Reichman, J.H., & Stiglitz, J.E. (eds.). (2014). *Intellectual property rights: Legal and economic challenges for development*. Oxford University Press.

Corredoira, R.A., & Banerjee, P.M. (2015). Measuring patent's influence on technological evolution: A study of knowledge spanning and subsequent inventive activity. *Research Policy*, *44*(2), 508-521.

Dawson, P., & Daniel, L. (2010). Understanding social innovation: A provisional framework. *International Journal of Technology Management*, *51*(1), 9-21.

Defourny, J., & Develtere, P. (1999). Origines et contours de l'économie sociale au Nord et au Sud. *L'économie sociale au Nord et au Sud*, 25-50.

De Jong, J.P., von Hippel, E., Gault, F., Kuusisto, J., & Raasch, C. (2015). Market failure in the diffusion of consumer-developed innovations: Patterns in Finland. *Research Policy*, *44*(10), 1856-1865.

Deng, Q., Hine, M.J., Ji, S., & Sur, S. (2019). Inside the black box of dictionary building for text analytics: A design science approach. *Journal of International Technology and Information Management*, *27*(3), 119-159.

Desmarchelier, B., Djellal, F., & Gallouj, F. (2020). Mapping social innovation networks: Knowledge intensive social services as systems builders. *Technological Forecasting and Social Change*, *157*, 120068.

Dosi, G., Marengo, L., & Pasquali, C. (2006). How much should society fuel the greed of innovators? On the relations between appropriability, opportunities and rates of innovation. *Research Policy*, *35*(8), 1110-1121.

European Union (2013). *Communication from the Commission, Europe 2020–A strategy for smart, sustainable and inclusive growth*. COM (2010) 2020 final 3 March 2010.

Gallini, N., & Scotchmer, S. (2002). Intellectual property: When is it the best incentive system? In *Innovation policy and the economy* (pp. 51-77). National Bureau of Economic Research.

Gerken, J.M., & Moehrle, M.G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, *91*(3), 645-670.

Grimmer, J., Roberts, M.E., & Stewart, B.M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hall, B.H., Jaffe, A.B., & Trajtenberg, M. (2001). The NBER patent citations data file: Lessons, insights and methodological tools. *NBER working paper*, *8498*(3094), 1-74.

Hargadon, A. (2003). *How breakthroughs happen: The surprising truth about how companies innovate*. Harvard Business Press.

Hazel, K.L., & Onaga, E. (2003). Experimental social innovation and dissemination: The promise and its delivery. *American Journal of Community Psychology*, *32*(3-4), 285-294.

Henry, C., & Stiglitz, J.E. (2010). Intellectual property, dissemination of innovation and sustainable development. *Global Policy*, *1*(3), 237-251.

Jeppesen, L.B. (2021). Social movements and free innovation. *Research Policy*, *50*(6), 104238.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435-1457.

Kuhn, J.M., & Thompson, N.C. (2019). How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, *26*(1), 5-38.

Kuhn, J., Younge, K., & Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, *51*(1), 109-132.

Lee, P. (2014). Social innovation. *Washington University Law Review*, *92*(1), 1-72.

Lemley, M.A. (2013). Software patents and the return of functional claiming. *Wisconsin Law Review*, 905-964.

Lengyel, A., & Botta-Dukát, Z. (2019). Silhouette width using generalized mean—A flexible method for assessing clustering efficiency. *Ecology and evolution*, *9*(23), 13231-13243.

Lerner, J., & Seru, A. (2017). The use and misuse of patent data: Issues for corporate finance and beyond. *NBER working paper, 24053, 1-92*.

Markard, J., Raven, R., & Truffer, B. (2012). Sustainability transitions: An emerging field of research and its prospects. *Research Policy*, *41*(6), 955-967.

Miric, M., Jia, N., & Huang, K.G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, *44*(2), 491-519.

Mitkov, R. (ed.). (2004). *The Oxford handbook of computational linguistics*. Oxford University Press.

Montgomery, A.W., Dacin, P.A., & Dacin, M.T. (2012). Collective social entrepreneurship: Collaboratively shaping social good. *Journal of Business Ethics*, *111*(3), 375-388.

Moulaert, F., Martinelli, F., Swyngedouw, E., & Gonzalez, S. (2005). Towards alternative model (s) of local innovation. *Urban Studies*, *42*(11), 1969-1990.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management*, *20*(4), 903-931.

Moser, P. (2013). Patents and innovation: Evidence from economic history. *Journal of Economic Perspectives*, *27*(1), 23-44.

Mulgan, G., Tucker, S., Ali, R., & Sanders, B. (2007). *Social Innovation: What it is, why it matters, how it can be accelerated*. The Young Foundation.

Mumford, M.D. (2002). Social innovation: Ten cases from Benjamin Franklin. *Creativity Research journal*, *14*(2), 253-266.

Nameroff, T.J., Garant, R.J., & Albert, M.B. (2004). Adoption of green chemistry: An analysis based on US patents. *Research Policy*, *33*(6-7), 959-974.

Nelson, R.R., & Sampat, B.N. (2001). Making sense of institutions as a factor shaping economic performance. *Journal of Economic Behavior & Organization*, *44*(1), 31-54.

Nicholls, A., & Murdock, A. (eds.). (2011). *Social innovation: Blurring boundaries to reconfigure markets*. Palgrave Macmillan.

Pel, B., Haxeltine, A., Avelino, F., Dumitru, A., Kemp, R., Bauler, T., ... & Jørgensen, M.S. (2020). Towards a theory of transformative social innovation: A relational framework and 12 propositions. *Research Policy*, *49*(8), 104080.

Pol, E., & Ville, S. (2009). Social innovation: Buzz word or enduring term? *The Journal of Socio-economics*, *38*(6), 878-885.

Teodoridis, F., Lu, J., & Furman, J. L. (2020). *Measuring the direction of innovation: Frontier tools in unassisted machine learning*. Available at SSRN 3596233.

van der Have, R.P., & Rubalcaba, L. (2016). Social innovation research: An emerging area of innovation studies? *Research Policy*, *45*(9), 1923-1935.

Von Graevenitz, G., Graham, S., & Myers, A. (2021). Distance (still) hampers diffusion of innovations. *Regional Studies*, https://doi.org/10.1080/00343404.2021.1918334.

Wall Street Journal (2019). *Big Business and its 'Stakeholders'*. 19/8/2019

Ziedonis, R.H. (2008). On the apparent failure of patents: A response to Bessen and Meurer. *Academy of Management Perspectives*, *22*(4), 21-29.

Tables of results

TABLE 1. Descriptive statistics: Spanish inventors

Variable	obs.	mean	std. dev.	min	max
Spanish inventors (count)	20288	2.198	1.731	1	20

Tabulation

Spanish inventors (count)	Freq.	Percent	Cum.
1	10207	50.31	50.31
2	3820	18.83	69.14
3	2743	13.52	82.66
4	1538	7.58	90.24
5	933	4.60	94.84
6	450	2.22	97.06
7	277	1.37	98.42
8	133	0.66	99.08
9	78	0.38	99.46
10	47	0.23	99.69
11	18	0.09	99.78
12	12	0.06	99.84
13	10	0.05	99.89
14	7	0.03	99.93
15	6	0.03	99.96
16	1	0.00	99.96
17	1	0.00	99.97
18	2	0.01	99.98
19	3	0.01	99.99
20	2	0.01	100.00
Total	20288	100.00	

TABLE 2. Word clusters

cluster 1	cluster 2	cluster 3		cluster 4
aquaculture	agrobiodiversity	development	security	economic
biodiversity	assistive	nation	fraud	indigenous
biomass	biofuels	inclusion	expression	academic
sustainability	biomimetics	abortion	gender	energy
carbon	biodegradable	abuse	genocide	fuel
abatement	bioeconomy	activism	partnership	water
sequestration	cap-and-trade	advocacy	globalization	certified
climate	carpools	bullying	governance	organic
climate-related	cleantech	campus	movement	civil
coal	cogeneration	capacity	destruction	neutral
gardens	composting	pricing	health	clean
desertification	desalination	economy	healthcare	sanitation
conservation	ecologically	rights	identity	coastal
ecological	ecotourism	freedoms	immigration	collective
ecology	renewable	child	migration	comparative
ecosystems	environmentally	labour	impact	consumer
environmental	ergonomics	childhood	investing	corporate
fishery	fair-trade	citizenship	inclusivity	crosscultural
forestry	fuel-free	society	city	cultural
fossil	greenfield	change	inequality	diverse
freshwater	landfill	code	innovation	elderly
forest	liveability	action	integration	emerging
stewardship	locally-sourced	colonization	unions	ethical
habitat	low-emitting	imperialism	wage	ethnic
hydroelectric	microcredit	education	marginalization	ethnographic
remediation	microfinance	community	masculinity	food
ocean	micro-lending	engagement	minorities	foreign
pollution	nitrification	participation	communities	aid
alleviation	non-timber	service	misrepresentation	gay
reclamation	reprocessing	protection	dilemma	global
regeneration	paperless	consumption	obligations	mental
agriculture	photovoltaic	philanthropy	morality	public
tourism	postconsumer	responsibility	multiculturalism	human
toxicity	recycled	corruption	safety	humanitarian
farming	recycle	perception	peace	hybrid
wetlands	rechargeable	diversity	privacy	inner
	recyclable	literacy	accountability	international
	research-and-development	democracy	knowledge	gas
	retrofit	diaspora	race	lgbtq
	soy-based	disability	refugees	local
	sustainable	discrimination	religion	low-income
	value-creation	inequity	enterprise	marginal
	systems-based	justice	wellbeing	mass
	triple-bottom-line	efficiency	stakeholder	moral
	value-added	elitism	initiative	natural
	vanpools	economies	growth	ngos
	vegetable-based	employment	fairness	nonprofit
	wastewater	equity	violence	nuclear
		empowerment	volunteerism	religious
		resources	welfare	responsible
		savings	workplace	rural
		regulation		sexual
		resilience		reproductive
		orientation		social
		activist		socially
		capital		socioeconomic
		cohesion		solar
		environment		special
		assessment		need
		equality		transport
		responsibilities		third-world
		ethics		traditional
		morals		transportation
		poverty		standard
		exceptionality		waste
		feminism		

TABLE 3. Descriptive statistics: word counts

variable	mean	std. dev.	min	max
word count	66.213	101.568	0	2501
doc size	9283.457	11095.75	730	212734
ratio total	0.008	0.014	0	0.097
count abstract	1.034	1.881	0	21
size abstract	121.571	69.36	3	2907
ratio abstract	0.009	0.015	0	0.15
count description	54.331	80.213	0	1056
size description	7635.964	10330.169	519	212067
ratio description	0.008	0.007	0	0.083
count claims	10.848	36.594	0	1758
size claims	1525.922	1835.61	5	49803
ratio claims	0.008	0.019	0	0.77
ratio cluster 1	0.015	0.102	0	0.994
ratio cluster 2	0.013	0.097	0	0.996
ratio cluster 3	0.006	0.05	0	0.994
ratio cluster 3	0.005	0.022	0	0.868

Patents with highest word count

patent	ratio total	title
9019954	.097899	*Methods and apparatuses for handling public identities in an internet protocol multimedia subsystem network*
5348655	.07764	*Method for increasing the capacity of sewage treatment plant*
8515492	.069935	*Energy managed service provided by a base station*
8655325	.067616	*Provision of public service identities*
9140241	.066165	*Manageable hybrid plant using photovoltaic and solar thermal technology and associated operating method*

TABLE 4. Word clouds

abstract

description

claims

TABLE 5. Main models

Variables	1	2	3
word count	0.00108***	0.00156***	0.00162***
	(0.00040)	(0.00035)	(0.00033)
doc size	-0.00001**	0.00002***	0.00001***
	(0.00000)	(0.00000)	(0.00000)
backward citations			0.00724***
			(0.00120)
non-patent citations			-0.00024
			(0.00073)
inventors			0.06079***
			(0.01388)
large entity			0.32738***
			(0.06023)
constant	1.76985***	2.70132***	2.37489*
	(0.04039)	(0.49186)	(1.40919)
class dummies	no	yes	yes
grant year dummies	no	yes	yes
obs.	3846	3846	3640

Variables	4	5	6
word count abstract	0.03529**		
	(0.01480)		
doc size abstract	0.00022		
	(0.00046)		
word count description		0.00228***	
		(0.00046)	
doc size description		0.00001**	
		(0.00000)	
word count claims			0.00285***
			(0.00102)
doc size claims			0.00004***
			(0.00002)
backward citations	0.00731***	0.00730***	0.00729***
	(0.00123)	(0.00120)	(0.00121)
non-patent citations	0.00059	-0.00033	0.00052
	(0.00068)	(0.00074)	(0.00068)
inventors	0.07959***	0.05885***	0.07899***
	(0.01385)	(0.01393)	(0.01382)
large entity	0.28877***	0.31977***	0.31006***
	(0.06066)	(0.06033)	(0.06057)
constant	2.32413	2.38751*	2.36166*
	(1.42990)	(1.41238)	(1.42068)
class dummies	yes	yes	yes
grant year dummies	yes	yes	yes
obs.	3640	3640	3640

Negative binomial regression. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1


funcas