DATA SCIENCE JOURNAL

Giovani, B 2017 Open Data for Research and Strategic Monitoring in the Pharmaceutical and Biotech Industry. *Data Science Journal,* 16: 18, pp. 1–9, DOI: https://doi.org/10.5334/dsj-2017-018

PRACTICE PAPER

Open Data for Research and Strategic Monitoring in the Pharmaceutical and Biotech Industry

Baldissera Giovani
EU-Target Consulting, FR
info@eu-targetconsulting.com

Open data is considered the new oil. As oil can be used to produce fertilisers, pesticides, lubricants, plastics and many other derivatives, so data is considered the commodity to use and re-use to create value.

The number of initiatives supporting free access to data has increased in the last years and open data is becoming the norm in the public sector; the approach empowers stakeholders and nurtures the economy.

Even if at early stage, private companies also are adapting to the open data market. A survey was conducted to which thirteen companies of different size (from micro enterprises to world-leading pharmas) in the pharmaceutical and biotech sector and representing four business models archetypes of companies exploiting open data (aggregators, developers, enrichers and enablers) participated.

The information collected provides a snapshot of the use of open data by the pharmaceutical and biotech industry in 2015–2016. The companies interviewed use open data to complement proprietary data for research purposes, to implement licensing-in/licensing-out strategies, to map partnerships and connections among players or to identify key expertise and hire staff.

Pharmaceutical and biotech companies have made of the protection of knowledge a dogma at the foundation of their business models, but using and contributing to the open data movement may change their approach to intellectual property and innovation.

Keywords: open data; pharmaceutical; biotech; research; strategic monitoring; intellectual property

Introduction

Since its origins that can be traced back to the open source software movements, the idea of open data has walked a long way. From a 'niche' philosophy this openness has embraced different fields, and it has even entered in citizens' everyday life.

The reasons for this rapid adoption can be linked to a number of factors, one of these being the economic crisis and the need for new stimuli to the world economy. The World Bank declared "the combination of geographic, budget, demographic, services, education and other data, publicly available in an open format on the web, promises to improve services as well as create future economic growth" (Hogge, 2010). This is without counting more technical (e.g. research) types of data.

Important research initiatives in the sector of human health have been shaped as 'open' from the beginning. The Human Genome Project was an ambitious research project started in 1990 whose goal was to sequence the whole euchromatic human genome and produce information that could benefit many different fields, such as anthropology, bioinformatics, forensic science, genetics, molecular medicine, pharmacology etc. The Bermuda principles (1996) set the rules for the exploitation of all the data produced within the project, requiring that all the genomic sequences unveiled should have been freely available to encourage research and development and to maximize the benefit to society" (Smith and Carrano, 1996).

The pharmaceutical and biotech industry has traditionally lagged behind other industries for what it concerns the contribution to the open data movement. There are a number of reasons to that, the main one being the adoption of a 'closed innovation' approach (Chesbrough, 2003): knowledge and innovation as a core activity should never be shared or sold (Mascarenhas et al., 1998). This allows that knowledge and information is formalised for the companies to appropriate returns from R&D and convert it in intangible assets (patents) that are at the foundation of their monotonous (and so far successful) business models (Levin et al., 1987).

A number of reports and articles are available on open data, but to our knowledge no study has been undertaken so far for the pharmaceutical and biotech sector. As for oil, data has no intrinsic value, but it is what someone can do with it that makes data precious. A comprehensive study of the consultancy firm Deloitte (2012) described five business models 'archetypes' of companies exploiting open data:

- Suppliers: organisations that publish their data via an open interface to allow others to use and reuse it.
- Aggregators: organisations that collect and aggregate open data and, sometimes, other proprietary data, typically on a particular sectorial theme, find correlations, identify efficiencies or visualise complex relationships.
- Developers: organisations and software entrepreneurs that design, build and sell web-based, tablet or smartphone applications for individual consumption.
- Enrichers: organisations (typically larger, established businesses) that use open data to enhance their existing products and services through better insight.
- Enablers: organisations that facilitate the supply or use of open data, such as the competition website Innocentive, but are not themselves users or re-users of open data.

These business models are all based on different exploitation/use/re-use of open data to develop new products, tools and services with high added value that can be sold. It is possible to create value from data more directly? Data is information and different types of data (and their metadata) could be used for strategic monitoring. This study also collects information on the intellectual property strategies of pharmaceutical and biotech companies that use open data to develop their products and services: how do companies deal with the apparent contradiction of exploiting open data and protecting their businesses?

Materials and Methods

An online survey (Google form, see annex I) was distributed:

- via email to various biotech and pharmaceutical companies; targeting specific networks and associations e.g. the Pistoia Alliance (http://www.pistoiaalliance.org), the Yale Open Data Project (YODA, http://yoda.yale.edu), the eTricks network (https://www.etriks.org), the TranSMART foundation (http://transmartfoundation.org); and to personal contact points.
- using professional media, in particular Linkedin. The following groups were targeted: Open data research network (1 451 members), Professionals in the pharmaceutical and biotech industry (207 446 members), Biotech and pharma professionals network (91 547 members), France biotech: entrepreneurs in life sciences (980 members), Biotechnology and business France (3 821 members).

In a number of cases, selective (face to face or phone) interviews were organised.

Information was collected between January and May 2016 and data analysed in June 2016.

The questionnaire was structured in 4 sections. A general section meant to collect information on the interviewee: the company, its business and the general approach to open data (both in terms of using open data and contributing to the movement as a 'donor' of data). Section 2 and 3 focussed on the use of open data, in particular open data for research activities or open data for strategic monitoring. The last section (4) allowed to collect information on the exploitation of open data, its advantages and disadvantaged and the impact of open data on intellectual property strategies.

The number of open questions was reduced to the minimum and multiple choices were preferred in order to ease the treatment of data by statistical analysis.

Results and discussions

Representatives of thirteen pharmaceutical and biotech companies provided their availability to participate to the survey. Efforts were made in order to contact different types of companies, both in terms of size (annual turnover, staff size) and in terms of use of open data. For reasons of confidentiality, the identity of the respondents is not disclosed in this article.

Six companies interviewed are small companies having a staff of less than 25 and an annual turnover lower than 2.5 M€; three companies are medium-size organisations with a number of employees between 70 and 200 and revenues between 12 and 30 M€; four companies can be considered large companies, with staff up to 88,000 and revenues up to 60 B€ (**Table 1**).

No supplier (according to the categories identified in the seminal publication of the consulting company Deloitte, 2012) was represented. Four companies (company b, d, j and k) defined themselves as 'aggregators', that collect and aggregate (open) data to identify biomarkers and develop drugs, that develop molecular models for the design and optimization of drug candidates, that analyse and extract value from different datasets for clients in different fields (chemistry, cosmetics, pharmaceuticals, but also luxury, finance, etc.); one company (company i) falls under the category 'developer' as it commercialises apps based on algorithms developed and tested using (open) data. Six respondents (company c, e, f, g, h and m) answered as 'enrichers', and these were biotech and pharmaceutical companies exploiting open data to support internal R&D activities (drug/vaccine development, molecular and *in vitro* diagnostics, etc.). Two companies (company a and l) are 'enablerers', providing access and facilitating the exploitation to medical data (e.g. medical imaging data).

Using and contributing to open data

Eleven out of thirteen companies use open data to complement proprietary data. As the two companies that do not use open data for this purpose are the companies providing services to their clients (i.e. they are not interested to use research data themselves), we can consider that the use of open R&D data is common to all companies undertaking some kind of in-house research activity. The use of open data for strategic monitoring is less common, eight out of thirteen companies take advantage of open data to support business intelligence activities, either for themselves or on behalf of their clients. Despite the more restricted use of open data for this second activity, yet its importance is demonstrated by the fact that there are platforms developed as portals to facilitate the access to open data for their use in strategic monitoring, and a number of companies base their business on such an activity. As an example, the Company Expert System (previously Temis) developed Luxid® Navigator http://community.temis.com/home, a navigator that allows the exploitation of information from articles, patents, meeting proceedings and databases. As indicated in the company website, a number of (leading) pharmaceutical companies benefit of the semantic technology to support information management strategies.

If the companies interviewed use extensively open data, their attitude towards the sharing of internal data is more conservative (**Figure 1**). This is not surprising, pharmaceutical and biotech companies have inherited the business model of the chemical industry and patented pharmaceutical products account for 80 to 90 percent of sales (Friese et al., 2006). But these same companies are users of open data, we can then extrapolate that these data are originating mainly from activities of the public sector. A trend was identified.

Company ID code	Size	Staff	Annual turnover (€)
Company a	Micro	1	0
Company b	Medium	103	12 Millions
Company c	Large	50,000	25 Billions
Company d	Small	25	2.5 Millions
Company e	Micro	5	Not disclosed
Company f	Medium	68	30 Millions
Company g	Large	70,000	40 Billions
Company h	Large	1,500	500 Millions
Company i	Micro	5	500,000
Company j	Small	10	200,000
Company k	Medium	200	25 Millions
Company l	Small	10	250,000
Company m	Large	88,000	60 Billions

Table 1: Size of the companies surveyed.

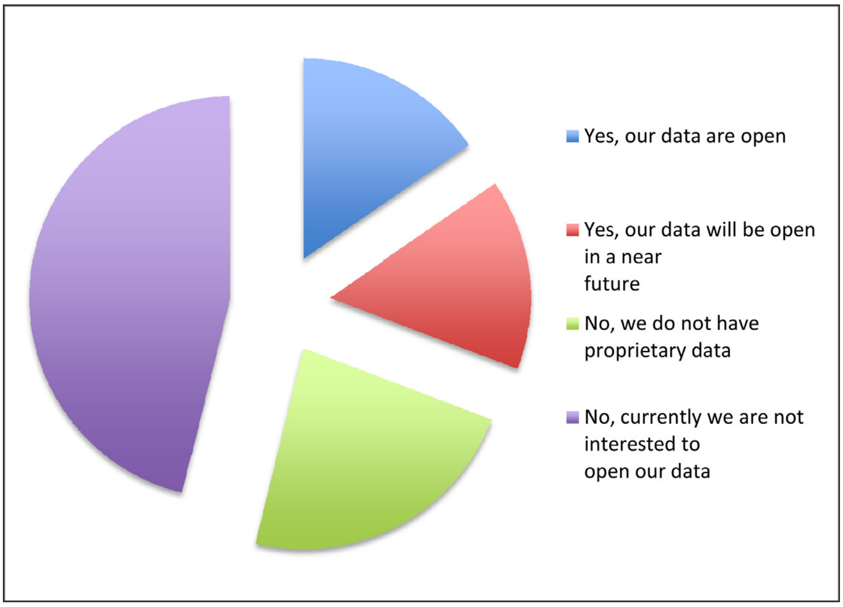

Figure 1: Openness of proprietary data. Yes, our data are open (15.5%); yes, our data will be open in a near future (15.5%); no, we do not have proprietary data (23%); no, currently we are not interested to open our data (46%).

The large pharmaceutical companies are those that either have some of their data open, or will open data in a near future. On the other side, the small and medium size companies do not share their data currently. Small-medium size companies are dependent on venture capital, and protecting knowledge is often the only collateral they can offer to investors. Moreover, opening data can have repercussions on the future exploitation of a product (validity of the patent), and the approach requires variegate skills and complex strategies more easily available within large companies, with the consequence that the smaller ones follow more classical IP approaches. Amongst the data made open by pharmaceutical companies molecular data (genomic, proteomic, metabolomic data) represents the large majority, because of the availability of well recognised infrastructures to host it; data on clinical trials records (World Health Organization International Clinical Trials Registry Platform and ClinicalTrials.gov registry[1]) is also open as since 1997 the Food and Drug Administration Modernization Act requires that data on the effectiveness of drugs clinical trials under clinical trials is made public. The adverse events reporting data[2] is another database that support the post-marketing surveillance on the safety of drugs or biological products.

Open research data to complement proprietary data

The use of open research data to complement proprietary data is linked to the use the companies make of the data as suppliers, aggregators, developers, enrichers or enablers (**Figure 2**). The aggregators surveyed have two different approaches for the use of open research data. Two of them (company b and d) only use less than 10% of open research data for their data analysis activities, while for the other two (company j and k) more than 80% of the data they handle is open data. No trend could be found, the pool of answers being too small. The enablers and the developer use less than 20% of open research data. This is surprising if we consider that these companies are small companies that do not conduct research activities themselves. The large majority of the data they use are non-open research data, which have the disadvantage that the companies have to pay to obtain them; on the other side this data is 'more reliable' than open research data, which is an essential aspect for companies that exploit the data to provide a service or develop a product. The amount of open research data that enrichers use varies from 10 to 80%. The differences are related to

[1] World Health Organization International Clinical Trials Registry Platform and ClinicalTrials.gov are registry and results databases of publicly and privately supported clinical studies of human participants conducted around the world.

[2] The US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is a database that contains information on adverse event and medication error reports submitted to FDA. The database is designed to support the FDA's post-marketing safety surveillance program for drug and therapeutic biologic products.

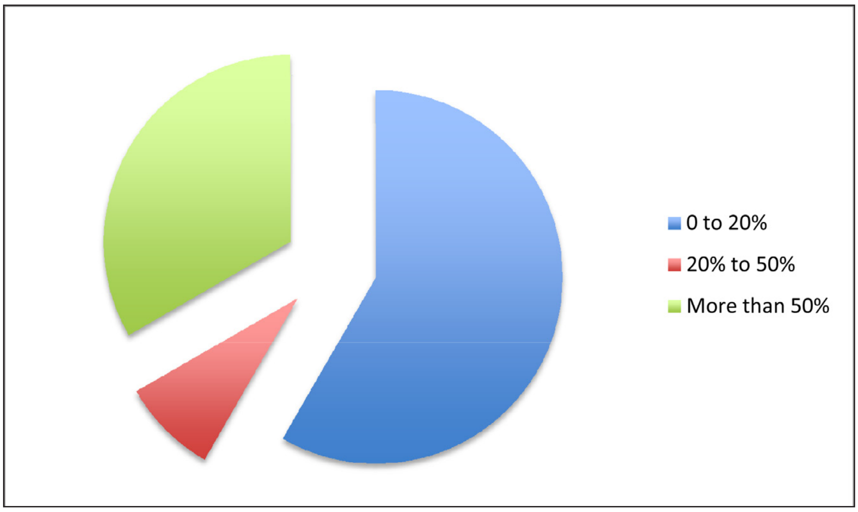

Figure 2: Use of open research data vs other sources of data. The percentages indicated correspond to the open research data. 59% of the respondents use open data in a limited way (blue); 33% of the respondents use more open data than other types of data (green); 8% of the respondents use equally different types of data (red).

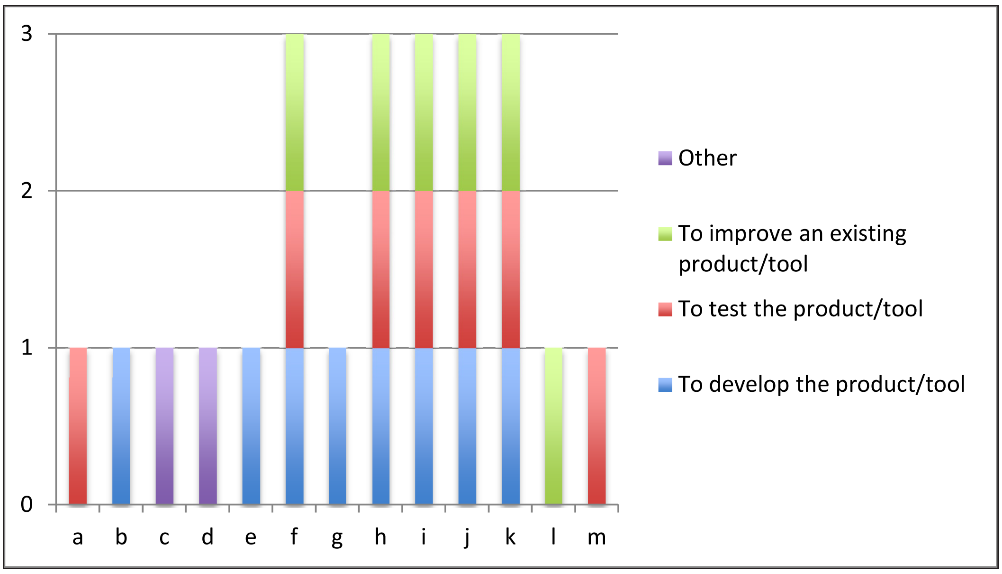

Figure 3: Use of open research data and Technology Readiness Level of the product. a, b, c, d, e, f, g, h, i, j, k, l and m are the identifiers of the companies interviewed (see Table 1).

the research activity carried-on and the availability of open research data in a specific field, but other factors like the strategy of the company (tightly linked to the companies' managers) probably play a role.

As indicated in the introduction, the value of data is linked to the use it is made of them. With research and development (R&D), pure research activities are covered, until development (pre-marketing) activities. The value of a product/service at different stages of development and the value of the research data linked to it varies with the 'maturity' of the product/service. **Figure 3** demonstrates that open research data is used to support activities over the entire Technology Readiness Level (TRL) scale.

Open data for strategic monitoring

Data from public research activities (in particular from the field of genetics and biochemistry) has been available to the entire scientific community for a long time, and this has contributed to defining the norm for the share and use of open research data. Other types of data are now becoming open in the pharmaceutical/biotech field, directly or indirectly linked to research data, that could be exploited for strategic monitoring. From the answers of the respondents it is clear that open data that is not used for pure research activities is mainly used to identify synergies and potential R&D collaborations. The use of data for the implementation of licensing-in/licensing-out strategies (by the intellectual property department) and for mapping partnerships and connections among players (by the marketing department) is comparable. The use of open data by human resources departments (e.g. to identify key expertise and hire staff) is the least common of the uses mentioned in the survey.

Currently, the classical tools used by companies for strategic monitoring are the patents and the publications; for their nature, patents provide information useful in the short term, while publications cover the medium and long-term (especially for the prospective ones). Most of this information is not available for free, as the access is subject to fees to the journals (in the case of articles) or to databases (for patents): Lexisnexis[3] or Questel[4] are just a few examples. Half of the respondents still take advantage of information from articles and patents (from 70% to 95%) compared to information from open data (from 5% to 30%). The other half already uses open data in an extensive way (from 40 to 100%) compared to other data sources (from 0% to 60%).

The availability of open data favours a new approach where data can be used for strategic monitoring. As demonstrated in this survey, this is already happening: the increased volumes of open data that will become accessible in future years and the development of tools for the textual and statistical data analysis, for data mapping etc. should contribute to the change.

When asked about their view on how important the contribution of open data could become for the strategic monitoring, the large majority of the respondents (10 over 13 respondents) considers that open data will become the main source of information.

Exploitation of open data, points of view

All the respondents use in a way or another open data in their activities. As such, they are best placed to identify benefits and disadvantages of such exploitation. The answers are very homogeneous. All point-out that the main advantages of open data are the fact that the information is rapidly accessible at virtually no price, compared to other types of data sources which have some sort of delay (e.g. publication of articles or patents). Moreover, a large amount of information (of any type) is available that has not been 'treated' (raw data) or averaged (individual data) which allows, for example, to develop more precise algorithms (taking into account all the data differences).

Open data shows a main disadvantage that has been identified by all the respondents: its variable quality, which means that users do not consider raw data reliable, thus preventing its exploitation. Other problems are the lack of standardization of data (what is shared, how it is presented, dependent by both the data producers and the infrastructures where data is hosted), the lack of data intentionality (for what purpose data was collected) and the lack of acquisition protocols (how data was produced), all essential information for the analysis of the data itself. The main consequence of all these gaps is that any user interested in exploiting open data must spend time and resources to analyse the datasets, homogenise them; in a word, curate them, before they can be used. Also, the success of open data could bring it to death; as more and different types of data become available through the internet, it becomes more and more complicated to find them and to exploit them in a way that takes advantage of their complexity and quantity.

The advantage of proprietary data, or published data, becomes then evident: they have been curated before they have been released; this reduces the type and the amount of data reaching the stakeholders, delays their availability but makes of the data trusted information.

What is the perception of the companies surveyed *vis-à-vis* the open data they use and their importance for the development of the product/service? Approximately half of the respondents (companies c, f, h, j, k and l) consider that open data is a key ingredient of their product, while for the other half (companies a, b, d, e, g, i and m) open data is a marginal ingredient of their product or service (**Figure 4**).

The intellectual property strategy for companies exploiting open data is of particular interest. The classical approach of companies was (and is) to own data and as a consequence of this ownership patents, copyrights or secret could be used to protect the products that are marketed. But the concept of open data is funded on a different logic. As a non-rival commodity the value of data is not in the data itself but in

[3] http://www.lexisnexis.com/en-us/gateway.page.

[4] http://www.questel.com.

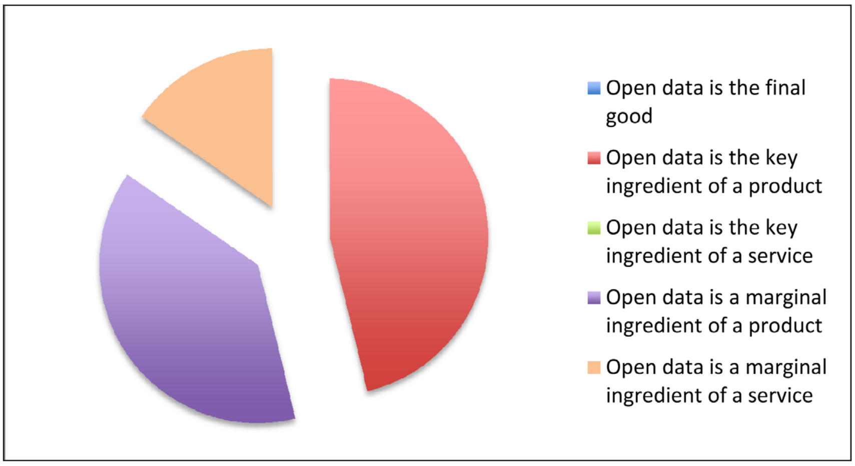

Figure 4: Value of the open data for the marketed product or service. Open data is a marginal ingredient of a service (15.5%), open data is a marginal ingredient of a product (38.5%), open data is the key ingredient of a product (46%).

its use and re-use: restricting the use of data has little sense. The business model of companies that use data produced by someone else for the development of their products is linked to the companies' ability to create value. Having this in mind, the fact that six out of thirteen respondents use the secret to protect their products is not surprising: the exploitation of data requires a know-how that is accessible to few, i.e. a technological barrier exists that, combined with the secret, will prevent competitors from exploiting the same area of business. Is the secret the best IP strategy? We estimate that in a rapidly changing environment where the trend is to disclose information and to collaborate and share knowledge and know-how (open data and open innovation), the secret is a short-term protection. But this is probably adapted to the short life of innovative products in the market. On the other side, open data does not prevent the use of patents or copyrights.

Questioned about their thoughts on the impact that open data could have on their IP strategy, most of the respondents (9 over 13) agree that open data will change the way IP is used by companies. If the aggregators (companies b, d and K) consistently answered that open data will not change their strategies (these are companies whose business stands on the exploitation of open data), the other companies (independently from their size and the use they make of open data) almost unanimously (with the exception of company f) agreed that open data will impact their IP strategy. Are we seeing here the premises of a more general change of the strategies of pharmaceutical/biotech companies to protect their processes and products?

Many different approaches (e.g. patents, copyrights, secret) are available and are efficiently used by companies in different fields. Mansfield (1986) theorised that manufacturing companies in most sectors (with the exception of the pharmaceutical sector) will continue to invest in R&D (i.e. will be able to protect their technology) even in absence of a patent system. Are we assisting to the end of this exception?

Conclusions

For the first time, a study was undertaken to collect information on the use that the pharmaceutical and biotech industry does of open data. If the number of participants to the survey is low and no statistically meaningful information could be deducted, yet the information received provides a snapshot of the approaches to open data of different (in terms of size, business, etc.) companies.

The answers collected demonstrate that even if at an early stage, the sector is preparing to use open data and benefit from it. It is known that open data is used to complement proprietary data for research purposes, yet this work demonstrates that open data can be used to implement licensing-in/licensing-out strategies, to map partnerships and connections among players or to identify key expertise and hire staff. According to the companies surveyed, open data will become the primary source of information for strategic monitoring, complementing patents that inherently provide the protection necessary to non-market coordination such as collaborations and alliances (Cohendet & Pénin, 2011). Besides the answers of the respondents, a number of companies have been identified that develop tools for clients to benefit of open data for their strategic monitoring.

We speculate that open data could change the approach to intellectual property: patentability requirements include novelty and non-obviousness, which are assessed from the state of the art. As patents currently represent 80% of the scientific and technical information worldwide (Burger-Helmchen et al., 2013) patent databases have been the best platform to find information on the state of the art that is not disseminated otherwise. Today, the amount of open data available to users increases over the time and the information is scattered and not easy to find (because of the variety of infrastructures and the way to access them worldwide); patent examiners will be confronted to the impossibility to know the prior art available and verify the non-obviousness postulate. This could lead to the same 'low-quality patent' controversy caused by the failure to screen applications against the prior art that resulted when the United States Patent and Trademark Office authorised the patentability of computer softwares (previously protected through copyrights) (Scotchmer, 2004). The logic solution is for patent offices not to rely on patents databases only (i.e. on known technology patents) but also on open data infrastructures to use the actual state of the art in their activities.

The same difficulties will be encountered by the patent applicants, as it could become more difficult for the patent holder to demonstrate novelty and non-obviousness in patent litigation.

Whatever their impact will be, open data will change companies' approach to intellectual property and innovation. But to let the revolution happen, efforts should be made to ease the access to open data, for example by reducing the number of entry points. National websites to upload/download/consult governmental open data already exist (e.g. https://www.data.gouv.fr/fr/, http://www.dati.gov.it, https://www.data.gov); by extending their remits they could play an important role as unique portals, data pools that should facilitate the accessibility of data and reduce uncertainties. The quality of open data also needs to be improved. New professionals figures (e.g. data managers) are arising, old ones are reinventing themselves (librarians) that can play a pivotal role to ensure that open data can be trusted and used. Still, standardisation of the data to reduce the variability that is intrinsic to the different sources of data is needed to ensure inter-operability; international coordination is the way forward that should span across countries, disciplines and sectors.

Additional File

The additional file for this article can be found as follows:

- **Annex 1.** Structure and content of the survey. DOI: https:/doi.org/10.5334/dsj-2017-018.s1

Acknowledgements

This work would not have been possible without all the people that have offered their time to answer to the survey. If most of them have decided to remain anonymous, they are the main contributors to this endeavour.

Competing Interests

The author has no competing interests to declare.

References

Burger-Helchem, T, Dintrich, A, Guittard, C and **Pénin, J** 2013 *L'innovation ouverte: définitions, pratiques et perspectives*. Chambre de commerce et d'industrie de Paris, Pénin, J (Ed.).

Chesbrough, H 2003 *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business School Press. DOI: https://doi.org/10.1108/14601060410565074

Cohendet, P and **Pénin, J** 2011 (December) Patents to exclude vs include: Rethinking the management of intellectual property rights in a knowledge-based economy. *Technology Innovation Management Review*, 12–17.

Deloitte 2012 Open growth. Stimulating demand for open data in the UK.

Friese, J, Jung, U, Röhm, T and **Spettmann, R** 2006 Intellectual property: an underestimated and undermanaged asset? *Journal of Business Chemistry*, 3(1): 42–48.

Hogge, B 2010 Open Data Study.

Levin, R C, Klevorick, K, Nelson, R R and **Winter, S** 1987 Appropriating the Returns from Industrial Research and Development. *Brooking Papers on Economic Activity*, 3: 783–820. DOI: https://doi.org/10.2307/2534454

Mansfield, E 1986 Patents and Innovation: An Empirical Study. *Management Science*, 32: 173–180. DOI: https://doi.org/10.1287/mnsc.32.2.173

Mascarenhas, B, Baveja, A and **Mamnoon, J** 1998 Dynamics of Core Competencies in Leading Multinational Companies. *California Management Review*, 40(4): 117–132. DOI: https://doi.org/10.2307/41165967

Scotchmer, S 2004 *Innovation and Incentives*. MIT Press.

Smith, D and **Carrano, A** 1996 International large-scale sequencing meeting. *Human Genome News*, 7(6): 19.

How to cite this article: Giovani, B 2017 Open Data for Research and Strategic Monitoring in the Pharmaceutical and Biotech Industry. *Data Science Journal,* 16: 18, pp. 1–9, DOI: https://doi.org/10.5334/dsj-2017-018

Submitted: 05 November 2016 **Accepted:** 15 March 2017 **Published:** 04 April 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS