Grandes bases de données comme facilitateurs de la créativité : le cas de l'industrie pharmaceutique.

MORATAL, Nuria

BETA, Université de Strasbourg

nmoratalferrando@unistra.fr

La biologie génère de très grandes quantités de données et les bases de données biologiques les collectent. Ces bases de données sont essentielles dans de nombreux domaines de la recherche en sciences de la vie, y compris la recherche médicale (Attwood et al., 2011). L'industrie pharmaceutique est confrontée à une crise de l'innovation depuis plus de 30 ans, malgré la croissance l'investissement dans des activités de R & D. Maintenant, ils s'appuient de plus en plus sur l'utilisation de bases de données et leurs recherches sont de plus en plus basées sur l'utilisation de techniques de Big Data et Intelligence Artificielle. Il y a plusieurs raisons de croire que ce changement va probablement s'accompagner d'une créativité accrue qui pourrait résoudre la crise de l'innovation dans le secteur. La recherche sur la créativité scientifique souligne l'importance de la diversité et de la distance des connaissances. Nous examinerons si l'utilisation de bases de données permet aux sociétés pharmaceutiques d'accéder à des connaissances plus distantes et variés mais aussi par quels moyens. Pour ce faire, nous utilisons une méthodologie d'étude de cas qualitative et interrogeons neuf sociétés pharmaceutiques qui utilisent les grandes bases de données publiques européennes EMBL-EBI. Nos résultats montrent que ces bases de données offrent en effet l'accès à une grande variété de données. Cette variété de données se présente sous la forme de variété dans son origine (principalement des données produites par d'autres communautés) et sur sa typologie (types de composés). Les bases de données permettent aussi d'établir de nouveaux liens entre des bases de connaissances. En ce qui concerne les types de résultats créatifs que l'on trouve, il s'agit des composés prometteurs et l'ouverture de nouvelles voies de recherche. Enfin, nous trouvons un résultat inattendu : les bases de données permettent la sérendipité.

Mots-clés : industrie pharmaceutique, créativité, distance de connaissance

Large Bio Medical Databases as facilitators OF creativity: the case of the Pharmaceutical Industry.

Biology is generating very large amounts of data. Today biological databases collect data produced by research in biology everywhere in the world. These databases are crucial in many fields of life science research, including medical research (Attwood et al., 2011). The Pharmaceutical industry has faced a crisis of innovation for more than 30 years, EVEN though the investment in R&D activities has not stopped increasing. Now they rely the more and more on the use of databases and their research has become data-driven. There are several reasons to believe that this shift is likely to come with an increase in creativity which would solve the innovation crisis of the industry. Research in scientific creativity highlights the importance of variety and knowledge distance. We will look at whether the use of databases allows pharmaceutical companies to access more distant knowledge. To do so we use qualitative case study methodology and interrogate nine pharmaceutical companies who use the Large European Public databases EMBL-EBI. The interviewees are all managers of bioinformatics departments of their companies. This position gives them a privileged view as they are the intermediaries between the researchers who use the databases daily and the databases themselves. Our findings can be summarized as follows. These databases are a crucial resource as they offer access to a high variety of data. This variety of data comes in the form of variety in its origin (mostly data produced by other communities) and on its typology (kinds of compounds and parameters associated). Additionally to this variety, databases allow scientists to make new connections between knowledge bases. Concerning the kinds of creative output that one kind find, it consists on one side on promising compounds and on the other side on the opening of new research questions and new research paths. Finally, we find an unexpected result which is that databases allow for serendipity, as it exposes a big amount of information to the entire scientific community. We develop all these aspects in the following lines to better answer to the research questions that were asked before.

Keywords: pharmaceutical industry, creativity, knowledge distance

1 INTRODUCTION

Science has always produced data. Since scientists started observing the world around them, they started taking notes on their observations and experiments where they would describe, measure and count. They have always as well preserved samples and specimens. Even before the existence of science as we know it today humans would take notes on lunar cycles, write about harvest cycles or write about important events for the community. These collections of notes are data and they have been accumulated for centuries and part of them is still used today. With each technical improvement, the quality and accuracy of observations was improved as well. More recently (from the 1950's) several instruments have been developed which allowed the massive collection of data. With the development of these instruments the quality of data has grown together with the size and number. The growth has been exponential and has led several domains of science to a change of paradigm as it has changed the ways scientists work and disciplines are conceived. It has been called the fourth paradigm change and it consists on science which is based on the exploitation of these big amounts of data (Hey et al., 2009).

Let us look at the case of biological databases. Biology is generating very large amounts of data. Managing this data has become a complex matter over time as it requires very large amounts of storage space and computing capacity (Attwood et al., 2011). This is why there is an increasing need of publicly available databases that analyse, integrate and summarize the available data, providing an invaluable resource for the biological community (Bolser et al., 2012). The data that is being produced is integrated constantly and we can observe a continuous increase in its size and use (Gong et al., 2011). The development of these biological databases has come with a growth on the use that medical research does of them. During the second half of the 20th century medical research had still an important part of randomized experiment and the few data used consisted of small in-house produced data silos. During the last decade the use of large datasets for health research including the case of pharmaceutical research has become the more and more extended and Data driven research has become highly important. The Pharmaceutical industry has faced a crisis of innovation for more than 30 years. Some argue that easy problems have been solved and that the current challenges of the sector are more complex and therefore scientific advance requires more time. Large companies could be investing more into fundamental science and pursuing long-term goals and challenges (Munos, 2009). In this paper we explore in which ways the swift of paradigm to a data driven research

can improve creativity and therefore the possibility to overcome that innovation crisis. To do so, this article is organized as follow. Section 2 builds up the theoretical framework that helps understanding which kinds of elements of the creative process we can expect to find for the case of databases. Section 3 presents the methodology that we use in order to fill those research gaps and complete and concretize the theoretical framework. We perform a qualitative case study with users of the EBI database. Section 4 presents the results of the case study and discusses them. We end with section 5 which will consist on a conclusion of the findings and a discussion of the perspectives for future work.

2 THEORETICAL FRAMEWORK

There is a growing body of literature aiming to understand scientific creativity. It is, however, still an understudied topic. When it comes to the specific case of databases and scientific creativity there is, to the best of our knowledge, no previous research that aims to understand that relationship. The objective of this section is to identify in the literature the concepts that are important for the understanding of the role of databases in scientific creativity. The concepts that we identify are used to bright up our understanding of the mechanisms, drivers and inputs potentially involved in the creative processes of science production. We focus particularly on those concepts that could apply to the case of biological databases and the use that the pharmaceutical industry does of them.

2.1 Findings on scientific creativity: the importance of variety

Knowledge production is the result of a combinatory process. Scientific discovery can be viewed as a form of human problem solving, a process which involves combination. Nelson and Winter (1982) said that “the creation of any sort of novelty in art, science or practical life – consists to a substantial extent of a recombination of conceptual and physical materials that were previously in existence.” In other words, to create new knowledge scientists use existing knowledge pieces and combines them. Following the same idea Arthur Koestler (1964) talks about bisociation which consists on combining two frames of thought (with concepts, ideas and perspectives) in order to build new creative knowledge. We study creativity as a process where several components are combined at different stages.

Creativity in science is defined as the creation of knowledge which is new and valuable. These two criteria are crucial for creativity and therefore they should both be always present when studying creativity. Novelty is involved during the earliest parts of the process, where the combinatory dynamics happen. Literature shows us that scientific creativity is strongly influenced by communication. When communicating strongly within an organization and across the borders of the organization scientists exchange knowledge, ideas and viewpoints and this has a positive effect on creativity. The reason is that they introduce diversity in their combinatory process and increase the chances of doing novel combinations. Following a similar logic, works on network brokerage argue that people who are placed at the intersection of heterogeneous social groups have an increased likelihood of drawing upon multiple knowledge sources, leading to the generation of new ideas. For example, managers who occupy brokerage positions are more often than others the source of good ideas (Burt, 2004; Rodan and Galunic, 2004). Diversity has been recognized as a driver of scientific creativity. Diversity comes in the form of different ideas and concepts and it can be applied to the kinds of knowledge that are put together during the process of science production. Researchers have shown how research that uses knowledge from a variety[1] of fields and involves scientists from a variety of backgrounds is more creative. Research organisations that allow for multidisciplinary research across departments, foster collaboration and promote mobility of researchers tend to be more creative (Heinze and Bauer, 2007; Hollingsworth, 2002; Zuckerman, 1967). Similarly, when teams include a variety of backgrounds or research is organized around problematics rather than disciplines, there is a tendency towards more creative outputs (Heinze et al., 2009; Lee et al., 2015). Because of their size and the integration of elements coming from all the subdisciplines of biology, one can expect Biological databases to introduce some of this variety of knowledge that can be used and combined.

Following Heinze and Bauer (2007), new knowledge appears in a diversity of forms.. Firstly, there is the formulation of new ideas or new sets of ideas that open a new cognitive framebring theoretical claims to a higher level of sophistication or challenge existing paradigms. An example of this kind of creative science is the Theory of specific relativity in physics by

[1] Diversity and variety are terms used to refer to the same idea in the literature on scientific creativity. They also appear as synonyms in the Collins English Dictionary. https://www.collinsdictionary.com

Einstein. In second place there is the discovery of a new empirical phenomenon that stimulates the building of new theory. A famous example of this kind of creativity would be how the observation of biodiversity led to the Theory of Evolution by Darwin. Thirdly there is the development of a new methodology. A new methodology, despite not being a scientific result by itself has the potential to solve theoretical problems that could not be empirically tested yet. Closely related to the previous one there is the invention of novel instruments that open up new research domains and new research questions that we could not imagine before. Finally there is the new synthesis of formerly dispersed knowledge. It consists of putting together ideas and connecting phenomena that were considered separately before, and putting them together into one same cognitive frame. It is at this last part of the production process of science that the notion of value appears. When performed in private organizations this new knowledge will need to prove its value. For instance a new promising compound in a pharmaceutical company will have to go through pre-clinical trials in order to prove it is useful.

2.2 Insights on knowledge distance and variety

As we have seen variety is crucial for the study of scientific creativity and the notion of knowledge distance very insightful and strongly related to variety. Science is done by recombining knowledge (Schilling and Green, 2011). The knowledge used when creating new knowledge can be more or less varied and this will have an impact on how creative the result is (Uzzi et al., 2013). The access to a variety of knowledge when recombining it to build new one is not easy. Let us look to public knowledge in general. The publication of knowledge doesn't make it accessible for the society. This knowledge must be found (the researcher needs to know it exist) and understood. Knowing this knowledge exist is already difficult and even when found it must be scanned, interpreted and learned. When exchanging knowledge with other individuals we find the same issue. It is not always easy for individuals to communicate when they have very different frames of thought. The ability to identify and understand external knowledge is absorptive capacity (Cohen and Levinthal 1990). Absorptive capacity refers to the capacity of an organization to acquire and assimilate knowledge coming from other organizations or individuals. The reason why absorptive capacity is needed is the existence of a distance between one's own (and an organization's own) knowledge and the knowledge that the individual wants to learn and use. Audretsch and Feldman, (1996) explain that the closest

the knowledge explored is to one's own knowledge; the easiest it will be to face the challenges of finding it and learning it. This is why through the process of knowledge production, actors often tend to use only knowledge that they are familiar with. It consists of knowledge that is in their own domain of expertise or in very related domains. Absorptive capacity helps firms to go further than their domain of expertise when looking for external knowledge. To explain the importance of Absorptive Capacity Nooteboom, (2000) uses the concept of cognitive distance, defined as "a difference in cognitive function which can be a difference in domain, range, or mapping". The bigger the difference in “mental schemas” between two or more individuals, the greater is the cognitive distance that separates them. This distance between their knowledge bases has the potential to create very novel connections that are valuable. As seen before, major discoveries come very often from combining a variety of disciplines and backgrounds. Nooteboom explains how greater absorptive capacity allows for a greater cognitive distance when using external knowledge.

Because large databases include a very large amount as well as a large variety of knowledge one can expect that they would make distant knowledge more accessible. The studies on the role of explicit knowledge in innovation justify this idea.

2.3 Insights on the role of explicit knowledge on creativity

We think that databases can have an important role in building absorptive capacity, and we show here the literature that suggests that. To the best of my knowledge, existing literature does not approach the impact that the extended use of Large Databases has on scientific creativity. For this reason, we explore the findings on the role of explicit knowledge on creativity as databases can be understood as a kind of explicit knowledge.

Explicit knowledge is codified and articulated as it is the case for databases. Explicit knowledge is also described as the one we formally learn at school and university, it is knowledge which can be expressed in words, numbers or equations and it is easy to share (Koskinen et al., 2003). This ability of being shared is what could give explicit knowledge a special role in creativity. Indeed explicit knowledge facilitates learning and therefore the acquisition of new knowledge and competences that, when combined with one’s own knowledge, have the potential to

increase creativity (Smith, 2001). One could thing that databases as well, because of some common aspects with explicit knowledge, can facilitate learning and therefore creativity.

However, explicit knowledge has often been described as the kind of knowledge that is produced and stocked at the end of the creative process and not as the type of knowledge that enables creativity. Explicit knowledge is related to organized tasks and routine. It assumes a predictable orchestrated environment. It is also aligned with a specific way of thinking which is logical, based on facts, that use proven methods and convergent thinking. This would mean that is not the kind of knowledge that enhances creativity as creativity requires divergent thinking and the use of uncommon concepts and ideas. Takeuchi and Nonaka, (1995) propose a set of basic patterns for creating knowledge in organizations. Within their framework explicit knowledge has an important role in learning and understanding which makes us insist on thinking that it could be crucial for creativity. For Takeuchi and Nonaka, (1995) as far as we are in a process of transforming explicit knowledge into tacit knowledge or tacit knowledge into explicit one there might be creativity. We propose, however, that the use of explicit knowledge is, by itself, relevant for creativity. Because knowledge is built up into previously acquired knowledge making knowledge explicit allows for a better understanding of this knowledge and therefore for being able to use it. For instance, Bourmand (1999) focuses on learning processes and shows that innovation is generated by the interactive process of learning where tacit and explicit knowledge are combined. We can therefore imagine that databases, as explicit knowledge, are easy to use and understand and therefore combination is made easier. If we look at the literature on knowledge distance, we see that understanding knowledge with which we are not familiar, can be an important challenge for firms and we think that databases could ease this challenge.

2.4 Research gap

Existing literature on scientific creativity throughs some light into which the role of biological databases on scientific creativity can be. It leaves, as well, some open questions and highlights some gaps in the literature that we aim to fulfil.

We see, in first place, how variety in the knowledge that is combined is important for scientific creativity. We have some reasons to imagine that the growth of the use of biological databases

has brought a growth in the variety of knowledge that is used. Biological databases integrated several kinds of compounds as well as a multitude of parameters associated to these compounds. One can easily imagine that this brings a variety of knowledge that is higher than the one companies had when working exclusively with in-house produced data. We don't know, however, the extent of this effect and how can it precisely affect novelty in the combination process. It is for this reason that we will investigate in first place how can databases bring variety and novelty to the combinatory process of science production.

We see, in second place, how it is important to use knowledge which is distant to an organization's own knowledge. In order to use distant knowledge, organizations need absorptive capacity. Databases could have a positive effect on absorptive capacity, and therefore on creativity, by facilitating learning. Indeed, the literature shows how explicit knowledge (a category that databases fit into) can facilitate learning and hence creativity. We will explore whether this happens, and which mechanisms are involved in this part of the process.

3 PRESENTATION OF THE CASE STUDY, EBI DATABASES AND METHODOLOGICAL APPROACH

To empirically explore the impact that databases have on scientific creativity and the suitability of our theoretical framework, we consider the case of the European Bioinformatics Institute (EBI). EBI is one of the main providers of public biological databases. We use a qualitative case study methodology with semi directed interviews with key actors. Case study methodology is recommended when looking at processes and trying to understand and explain complex phenomena (Yin 1994, Eisenhardt 1989 and 2007). It is also recommended when doing exploration of an understudied topic. Although we are not doing pure exploration and we do not use grounded theory, our approach nevertheless includes an exploratory dimension. Below we will see the details on the history of EBI and the nature of the databases that it provides, the details on the profile of the interviewees and the reason they were chosen and finally the strategy to be able through interviews to answer our research question.

3.1 Description of the EBI databases

The databases chosen for this research are the ones offered by EMBL-EBI, the European Bioinformatics Institute. EBI is a research institute but also a provider of databases containing biological compounds such as proteins or genes information. This database is the largest bioinformatics resource provider in Europe and is used all around the world. The access to these databases is completely unrestricted and free. The users of these resources are both public and private researchers. Our research focuses on the case of Pharmaceutical companies.

EBI origins lie in the first Nucleotide Sequence Data base that was established in 1980 at EMBL in Heildelberg, Germany. The initial goal was to create a central database of DNA sequences submitted to academic journals. It began with very modest aspiration of simply abstracting information from literature but soon it started receiving data directly. Universities and laboratories all over the word upload the results of their experiments and observations directly into the EBI databases. This reduces the job to EBI to that of verification. It poses however some challenges as the amount of data being inserted is growing at high speed. In addition, the magnitude of the database grew in scale when the Human Genome Project finished in 2003 and all the produced data had to be integrated. It was an international scientific research project with the goal of determining the sequence of all the three billion nucleotide base pairs that make up human DNA, and of identifying and mapping more than 100.000 genes of the human genome from both a physical and a functional standpoint. It is still today the world's largest collaborative biological project. EMBL played an important role on this project and when EBI database was established it integrated rapidly all the results from the Human Genome Project. This gave EBI more visibility and therefore more use and more popularity. Additionally to the accumulation of data produced by other institutions, and offering them, EBI employs scientists and produces research of its own, exploiting the database as well as producing data and introducing them.

EMBL-EBI started with two databases, one on nucleotide sequences and another one for protein structure but with time it has diversified, and it provides now resources in all the major molecular domains. It gives access to freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry. The services entail not only data

archiving but also data curation and integration. They allow users to query EBI's large biological databases programmatically, eventually to build data analysis pipelines or to integrate public data with users' own applications. The six core data resources are operated by relatively large teams of 15 to 20 people (scientific curators, software engineers, bioinformaticians, and visitors including PhD students). [2]

3.2 Research Strategy

In order to have an answer to our research questions, the users from the pharmaceutical sector were interrogated directly about the use of these databases. The specific target was the heads of Bioinformatics departments in big Pharmaceutical companies. Why this specific sector and why only big companies? As we have seen there has been an innovation crisis in the pharmaceutical Industry and it is interesting to ask whether the change on ways of working within this sector is likely to bring some creativity and, in the end, innovation. The chosen companies belong to what is commonly known as "big pharma" which are very large pharmaceutical companies. The reason to focus on Big Pharma and not small companies is because it is precisely within big pharma that there has been an innovation crisis. The pharmaceutical market is dualized, with a few very large companies and a big number of start-ups. Within start-ups innovation is the key to their survival, and we can expect, therefore, that creativity will be present. For Large companies, however, where the strategy lies more on the exploitation of well-established products, being creative is more challenging (Orsenigo and Malerba, 2015). Finally, an important reason for the choice is that the Pharmaceutical industry is concentrated; most of the market is controlled by a reduced number of firms and it is therefore relevant to focus particularly on them.

The first step, before the design of the questions and topics to be discussed during the interviews, was to learn about the context. The aim of this was to better understand the past and present use of data in the pharmaceutical companies as well as which kind of data are used and how are they used. This was done by means of desk research. The kinds of documents studied during this part of the processes are some policy reports concerning the provision and

[2] fhttp://www.ebi.ac.uk/

use of databases, OCDE and European Union reports on science, as well as the annual reports of the providers of those databases. There was as well a process of literature review with articles published in some of the mainstream scientific journals that discussed scientific policy. Table 1 shows a list of the most relevant documents used during this part of the process.

Table 1 Desk research: Institutional reports and scientific journals

Desk research
Institutional reports
Data Availability Policy Final report of Public consultation on Science 2.0, Open Science, European Commission.
Sharing data from large-scale biological research projects: a system of tripartite responsibility Excellent Science in the Digital Age, European Commission.
Scientific journalism and news
Computational Biologists: The Next Pharma Scientists? Science magazine. Michael Price April 13, 2012
An Explosion Of Bioinformatics Careers. Science magazine. Alaina G. Levine June 13, 2014
A decade's perspective on DNA sequencing technology. Nature Elaine R. Mardis 2011 470: 198-203
Science after the sequence. Nature News. Declan Butler reports June 2010

3.3 Choice of informers

The users studied participate in a Partnership Programme with EBI[3]. This partnership was created by EBI in order to keep informed of industrial users' needs. The participation in the partnership is open to any company that is ready to pay the participation fees. The participation on this Partnership and its impact on creativity are not studied in this article and for anonymity reasons we cannot disclose the names and companies of the interviewees. The reason to chose them was, firstly, that most of them come from what is commonly known ad "big pharma" and this kind of companies are the focus of our research. Secondly, the existence of this partnerships allowed us to meet them informally a few times during their quarterly meetings and build some

[3] This Partnership is The Industry Programme. It was used only as a way to identify industrial users from the Pharmaceutical Industry. The creativity that is the result from this partnership is not studied. The members of the Industry Programme may change from one year to another. https://www.ebi.ac.uk/industry

trust. This trust was crucial as the pharmaceutical sector has traditionally been very closed and secretive.

The interviewees are all managers of bioinformatics departments of their companies. This position gives them a privileged view as they are the intermediaries between the researchers who use the databases daily and the databases themselves. This means they are in daily contact with both the users' needs and the data and technical possibilities. The second reason for choosing these individuals is that they constitute a homogeneous profile and this can give us an in-depth view on the sector. Since the methodology used is a case study methodology, it is appropriated to focus on a specific profile of user and seek for an in-depth knowledge of it.

3.4 Interviews

The interviews were semi directed (see topics treated in the following section). There was a set of topics, but no specific questions were included in order to avoid influencing the answers of the interviewees. They were encouraged to talk freely about their vision on the effects of these databases on creativity. In total, 9 companies have been studied. All the interviewees are at the head of the bioinformatics departments of their respective companies. There were two waves of interviews as well as informal discussions during a 2-day meeting organized by EBI in between the two waves of interviews. The format was face to face interview (with one exception of a Skype interview). The reason why there where two waves of interviews is that, during the first wave, companies were interrogated mainly about the value dimension of creativity. Some questions concerning the novelty part of creativity were done already but they were mostly exploratory. Creativity consists on both, value and novelty. Although novelty often comes earlier in the production process of science, value is a dimension that is more often studied and therefore it is easier to identify as there are more widely accepted tools and parameters. This dimension is also easier to observe because it is present very often at the output level. The first wave of interviews allowed us to understand the use that is done of these databases, how do they intervene in the production level and at which points can we find the notions of variety and distance. These two notions are the ones that are related to novelty. The second wave of interviews was more focused on these dimensions of variety of knowledge distance and on trying to identify novelty.

The interviews lasted between 30 and 120 minutes. After analysing them a second informal encounter was set to discuss some missing and incomplete information. This encounter happened with the entire group during one of the Industry Programme meetings. Afterwards a second wave of interviews was set, with three of the already interviewed people. In this case interviews lasted longer, from 78 to 115 minutes. All this information is summarized in Table 2. All interviews were recorded and transcribed verbatim, under the conditions of anonymity and confidentiality of information. Anonymity conditions here implies not only not disclosing the name of the people and companies involved.

Table 2 Overview of interviews

Company pseudonym	Informal discussion	1st wave	2nd wave
Ph1	Yes	90 min	80 min
Ph2	Yes	90 min	115 min
Ph3	Yes	60 min	78 min
Ph4	Yes	65 min	No
Ph5	Yes	60 min	No
Ph6	No	100 min	No
Ph7	No	30 min	No
Ph8	No	50 min	No
Ph9	Yes	No	No

The first wave of interviews was conducted between March and April 2015. Most of them took place at EBI during an Industry Programme meeting. All the interviews started with the general information on the company (size, employees, size of R&D department). The subsequent topics focussed on the main goal of the interview: characterising at the level of each company the role of bioinformatic databases in general and of EBI's ones in particular. More precisely our questions concerned the kinds of databases they used other than EBI (for instance, some private databases or public databases from other providers). We asked as well about the way they used them (how many people in the company were concerned by its use, frequency of access, etc.). We also asked about the importance of these databases (i.e. dependence for R&D activities). The same questions were asked about bioinformatic tools related to the databases. Then the

final part of the interview concerned the evolution of the role of these databases during the last decade, in order to understand the changes, they possibly brought to the R&D processes of the companies, and/or their contributions to the efficiency of R&D activities. Table 3 summarizes all this information. In order to ensure that there had not been any misunderstanding and that what we had understood what we had been told during the interviews, we contacted the interviewees and asked for confirmation of agreements with our interview notes. This interaction was done via exchange of e-mails and was completed by December 2015.

Table 3 Content of first wave of interviews

Question	Issues explored
Presentation of the company.	General numbers (Number of employees, number of employees in R&D, number of employees in bioinformatics, annual revenue, R&D budget and bioinformatics budget
Which is the intensity and relevance of use of EBI and other databases?	Which databases are used, how often are they used, which is the level of integration between in-house databases and EBI databases and degree of dependence
Which is the intensity and relevance of use of bioinformatics tools in general?	Which kinds of related tools are needed, how often are they used, which is their role in the production process
Which have been the changes and improvements in the way research is done and the results you have?	How bioinformatics in general and EBI in particular have changed the way they work and the research possibilities, comparison to how science was made before and quantity of resources they have access
How did databases bring those improvements?	Exploration of several topics

After completing this first wave of interviews, during December of 2015 some informal conversations took place. We went to a workshop of the Industry Programme to discuss with the participants about our conclusions and go a bit more in depths on some topics that needed further discussion.

Table 4 Content of informal discussions

Topics discussed
Importance and dependence of today's pharmaceutical research on Databases
Evolution of the use, current challenges and future possibilities
The importance of spaces such as the Industry Programme for the community

The second wave of interviews consisted in the confirmation of the knowledge learned in the first wave of interviews as well as achieving a more detailed an articulated vision on how databases can facilitate the access to a higher variety of knowledge as well as how this variety impacts creativity. As we can see in Table 4, we asked specific questions which were answered with extensive detailed explanations. Firstly, we asked interviewees were asked about sources of novelty and types of distant knowledge. To go deeper into these questions' interviewees were asked whether this meant an access to knowledge coming from other communities, other disciplines and other methods. After these questions, we assessed the connecting property of databases. How do Large Biological Databases connect pieces of knowledge that remained unconnected before? Finally, creativity and kinds of creativity were discussed, as well as the possible risks that the dependence on databases might bring.

Table 5 Topics treated in 2nd wave of interviews

Question	Issues explored
Which is your perception of the creativity in R&D?	Kinds of creative process: introduction of new methodologies, new scientific paths or new fields of research
Do databases provide access to a higher variety of knowledge?	Databases provide access to a higher variety of pieces of knowledge that are introduced in the research. Including knowledge from other disciplines and communities.
Do they have an impact on the amount of disciplines and methods used?	The large amount and variety offered by the database requires from new hiring policy. Multidisciplinary teams and variety of methods.
Do databases connect distant knowledge bases?	Databases and bioinformatics facilitate the connection of pieces of knowledge that were difficult to connect otherwise
Thoughts on risks	Lock in effects, fashions and trends.

4 RESULTS

As explained before, the interviews have been written down as verbatim and analysed. Because we part from an abduction logic, the methodology used for the analysis is to ask questions to explore the topics identified by the literature review (variety, knowledge distance, novelty and value) and go from the general to the more specific in order to understand how databases can contribute to these different mechanisms that have traditionally been associated to creativity.

For the parts of the interviews that could not be associated to one of our large themes we attributed it new codes, that emerged fully from the fieldwork.

4.1 Analysis of transcript (first wave of interviews)

This first wave of interviews consisted on the exploration of five general topics. This exploration had, as a result, the emergence of the 13 important concepts. These five general topics are: dependency on EBI and public databases, the improvements that the use of these databases has allowed for, the changes in the way that science is performed, the origins of those changes and how these databases can be a source of variety and novelty in science.

Dependency on EBI and other public databases

When asking about the dependency of the research activity on EBI all the interviewees said that, although it is their preferred source of data, they would easily continue their normal research activity by using the databases offered by some of the alternative public providers of biological databases. When asked about the dependency on public databases in general (and not only EBI ones) we can observe a polarization in the answers. On side we have those interviewees that consider that the advancement of science requires the existence of public databases. The most common argument that we find here is that no private provider could offer the quantity and accuracy of the data offered by EBI and the other big public database providers. If all of these databases were to stop existing, the advancement of science would suffer greatly. Other interviewees, however, claim that although without large databases science could not be done as it is done today, their company could continue doing research. This research would simply be done in other ways. They most common argument is that it is only recently that they use these databases and not long time ago they could easily do without them, with in-house databases or small databases provided by private companies. Finally, all of them think that the dependency on public databases is growing and they have become the more and more important in their companies for the last years.

Improvements offered by the use of large EBI databases

The second topic of discussion consisted on the improvements offered by the use of public databases and more particularly, EBI ones. There is rather uniformity concerning the answers

of the different people that were interrogated. The improvements that EBI concern the quality and quantity of data that is available to do research. This has as a result an increase on the speeding of the research processes. With EBI the data they can access is considered to have better quality than the data offered by private companies but also more reliable than the data that used to be produced in-house. This increase in quality comes in the form of accuracy of the data, number of parameters and also reliability (it is less likely to find mistakes). Concerning the quantity, the amount of compounds that are available is simply very big and cannot be compared to those offered by private providers or to the databases that are built inside the company. These two characteristics have allowed companies to save up some time and speed up the production process.

Changes on the way science is done

Most of the interviewees agreed that the way they perform science today is determined by the existence of these databases. It has changed the way that problems are posed and solutions are changed. The projects that are proposed today depend on the existence of large biological databases and without them those research projects could not be performed. The kinds of research questions asked would simply be different without the access to this kinds of databases.

Origin of changes

The changes on the way that science is performed come, in first place, from the quantity and quality of the available data. We asked which other reasons have made databases become as important as they are today in the production process of science. Three main factors were highlighted: access to variety, aggregation and curation. The access to a variety of compounds emerges in the interviews by its own before asking specifically about it. Most interviewees explain how EBI databases allow them to identify and understand a wide variety of data (i.e. variety of compounds and variety of parameters available). They claim that to have this effective access is not enough by having the information available and open. They need this information to be efficiently aggregated and curated (which means that the accuracy is verified and that the parameters are correctly written). EBI databases offer this.

Sources of novelty and variety

Because variety is important and because the topic emerged easily and naturally in most of the interviews, we asked what kind of variety do these databases offer and more precisely which factors allow for this variety. Concerning the kind of variety available the answers were rather vague and they simply talked about data that comes from other organizations than their own and data that are different than the ones they usually produced. This vagueness came also from our lack of understanding of the subject and motivated the second wave of interviews. What we found when exploring this topic is that what allows the access to this variety of data are the standards and ontology tools that EBI offers together with their databases. EBI's databases use some standards to express the multiple information and parameters of the data. Thanks to this standards the data can easily be understood, even when they are very different from the data one is used to produce. Ontologies are language tools and their role on access to variety follows the same logic. One of the traditional problems that scientist faced in the area of biology to use external knowledge is the names of compounds, particularly proteins and genes. Because there are hundreds of thousands of proteins and genes there was, up to now, no a common way to name them. Because of this, when looking for information on a protein, for example, scientists needed to know the name used by different groups of scientists to name that same protein. Ontologies are language tools that allow overcoming that limitation and have been defined by several interviewees as the most relevant contribution of EBI databases.

4.2 Analysis of transcripts (2nd wave of interviews)

During the second wave of interviews, the interviewees were asked more specifically about creativity in general as well, the variety of data they can access, and which are the creative outputs they consider to have gained thanks to the use of publicly available databases. There were four questions that were asked from which a total of nine topics emerged. The questions asked consisted on: the impact of the use of databases on creativity, the types of variety that are allowed by the use of the databases, the possibility to connect distant knowledge and the kinds of creative output. There was as well the emergence of an unexpected topic: serendipity. The reason to ask first for creativity in general and later on go further into the detail on the variety of knowledge that is used, the use of distant knowledge and the creative output is the

difference between which is our theoretical conception of creativity and what do the interviewees understand as creativity.

The impact of the use of databases on creativity

When asked about creativity, the three people that were interviewed answered directly by asking what we consider as creativity, our definition was challenged but we always found some common ground for discussion. Most interviewees considered that EBI database help them to be more creative. Because of the existence of such a large amount of data, research is not randomized anymore and there is a need to understand the disease. Then they have to design algorithms that will explore the data and hopefully find some promising compounds. This way of working forces them to think in different ways which they consider to be creative. They also consider that the existence of teams with mixed backgrounds (for instance data scientists and biologist) forces them to put together different kinds of knowledge and different kinds of expertise and therefore it makes them more creative.

The types of variety

As seen during the first round of interviews, this second round confirms the fact that EBI databases provide access to a wider variety of data and that this has an impact on creativity. More specifically this variety is expressed in terms of variety in the origin of data (particularly data coming from a wider variety of disciplines, and from multiple disciplines) and typology. Additionally, this variety of data also imposes a varied background in the team and a multiplicity of methods in the firm to process them.

We were explained how; traditionally researchers in biology and pharma used only data coming from their close circles. This means for example data coming from the same community of researchers (i.e. community of cancer research, community of heart disease research, community of Alzheimer disease research, etc.). This was, as explained during the first wave of interviews, because the way they named things forbids them to identify the research that concerned some specific genes or proteins. Thanks to the existence of ontologies, researchers can use data coming from a variety of communities and origins. Indeed, not only the research community would determine how the data are named, the country or the laboratory of origin of the data had an impact as well. EBI allows as well to access to a wider variety in terms of

typology of data (i.e. protein function, protein expression, protein sequence, protein/gene interaction, etc.). This is allowed thanks to the existence of standards which permits researchers to understand easily the data and its parameters.

Additionally, this variety present in the databases themselves forces companies to have variety in their research team. Because of the multiple kinds of data and the complexity of the problems that this data can be used for, research teams require a variety of backgrounds. They need to have computer scientists, data scientists, biologists, experts on different kinds of compounds, etc. This variety of teams working together will bring a variety of methods, tools and expertise that will all be put together in order to solve complex problems.

The possibility to connect different knowledge

Interviewees explained how, these databases have allowed them not only to access a varied typology of data, but also to be able to make connections among these data. This means, for example, to put together protein structure data and protein function data and use algorithms that will explore both of them together.

The kinds of creative output

We asked the three interviewees which kind of creative output comes from the use of EBI databases. It is important to notice that here we are not yet talking about drugs; the output from the research process is at a much earlier stage than the drug development. Indeed, large biomedical databases play a role in the very early stage of research and the creative results consist in promising compounds new research questions that are not on the continuity of previous research, opening new scientific paths and the development of new methodologies. When using EBI the compounds found to be promising are more likely to be validated. This means that the promising compounds are more likely to continue to the stage of pre-clinical trials. A surprising result consisted in considering that not only they found promising compounds. Thanks to the use of EBI databases researchers can open new research paths and ask questions that they would not have thought about without the existence of the data.

Unexpected emerging topic: serendipity.

Finally, an unexpected result was the role that these databases have in terms serendipity or happy accident. Indeed, we found that the existence of a critical mass of people facing a critical mass of data is crucial for someone coming up with the right idea. This is something that came up in the first of the interviews performed and that was later confirmed, although with less enthusiasm, by the other two interviewees. Interviewees explain that, while facing the "full picture" with an "open mind" it is more likely to have "happy accidents". They all explained how the unexpected findings became more common as the use of databases spread. Asked for examples on this kind of effect the most common ones consisted on doing an unexpected finding while looking into a completely different research question. In one example, while trying to find a method to detect the DNA of a future child in a pregnant woman's blood, researchers found a method to detect cancer in people's DNA. It is, however, not the only type of serendipity. In another case, after months of trying to figure out how to identify the genes associated to a specific kind of cancer they found the solution in an unexpected way. Several algorithms were being used to target the genes that were thought likely to be associated with that certain kind of cancer. Simply by observing at some data and confronting it with their own knowledge, that group of researchers understood that they had to modify the algorithm they were using. That modification in the algorithm let them to find the genes associated to cancer.

4.3 Serendipity

Serendipity appears as an unexpected result. Creative science is often explained in terms of serendipity. We did not expect databases to have an impact on serendipity and therefore its theoretical implications were not studied earlie. We discuss here what literature tells us about serendipity and how to relate it to our empirical findings. Serendipity in science defines the notion of unexpected and beneficial discoveries. Historically it has played an important role in explaining how big breakthrough discoveries are made (Merton and Barber, 2004). The description which often refers to the notion of "happy accident" is rather imprecise and therefore a multiplicity of phenomena can be described as serendipitous. This is the reason why Yaqub (2016) has discussed the heterogeneity of the phenomena and created a typology of the different kinds of serendipity and the mechanisms that lead to them. By exploring the most famous cases of discovery through serendipity as well as the literature that studies them he created a typology.

His paper classifies serendipity according to how it shows up or in other words, what makes a discovery serendipitous. This leads to several kinds of serendipity, from which two are the most commonly cited by the literature. The first one, **Walponian** serendipity consists in targeted search leading to an unanticipated discovery. Researchers were looking into one problem and made a discovery into another. This happened sometimes as the result of an accident or simply because some effect (which wasn't the one they were looking for) showed up during the experiment. This is the most known kind of serendipity which led to the idea of a "happy accident". The second kind of serendipity is **Mertonian** and it occurs when targeted search solves a problem via an unexpected route. It consists in researchers trying to solve a problem and finding a solution in an unintended way. For instance, by accidentally mixing some components that were not expected to lead to a solution but do. These two kinds of serendipity correspond to the examples explained in the previous section of discoveries that were allowed thanks to the use of databases and were serendipitous.

In a second stage of his research Yaqub looks at the underlying mechanisms of serendipity. He shows how luck is important for serendipity to occur but it is a subordinate to researchers' skills, knowledge, curiosity and ability to share and communicate information. What they all have in common is the existence of an open-eyed or watchful state by the community. This is what makes researchers aware of what is out of the norm, which phenomena might be interesting. We can expect the use of databases to have an impact on serendipity by enhancing some the effects described here.

4.4 Discussion of results

In this section we want to answer the questions that came from the study of the literature and the identification the research gaps.

- **How do databases offer value to its users?**

We see that large public databases offer some value to companies that use them because they are changing the way they do science and they are the more and more dependent on large databases. Without them science could not be done in the way that is done today. The quality and quantity of the data is the reason for this choice. No private company could offer such an

amount of data and such quality in terms of accuracy and completeness. Having databases has allowed them to speed the production process.

- **Can databases have a positive effect on novelty? Do they have an effect on the variety of knowledge combined when performing science? How?**

The effective accessibility to this large amount of good quality data is allowed by its aggregation and curation. In other words, the fact that all the information is found together in the same platform and by using the same interface is a great facilitator of access. Additionally the accuracy and validity of this data is continuously verified and does not create problems to the companies that use them. The feature that is considered the more relevant from these databases by all the people interrogated is that they provide standards and ontologies, which are tools that allow the access to a wider variety of data. This variety of data comes in the form of variety in its origin (mostly data produced by other communities) and on its typology (kinds of compounds and parameters associated)

- **Can databases have a positive effect on novelty? Can they facilitate access to distant knowledge? How?**

The existence of standards and the aggregation of databases in one single interface also facilitate making connections between the different kinds of compounds that are find in the databases. This means that the variety found can be put together in a combinatory process and potentially has positive effects on creativity.

- **Which kind of creative output comes from the use of large public databases?**

Because we are at the very early stages of the drug production process the kind of output that we are facing is not, yet, a commercial one. Databases help, in first place, to find more promising compounds. Because of the big amounts of information concerning genes, proteins and other compounds, scientists can run algorithms that allow them to better predict which of this compounds are more likely to become a drug. Interviewees told us that, when using EBI databases, the targets identified (potentially drugable compounds) are more likely to be validated and go into the pre-clinical trials. A second surprising kind of creative output consists on being able to ask new research questions and open new research paths. This last aspect

comes from the fact of thinking about a possible question, a possible project, after being inspired by the data

- **Unexpected result**

An unexpected result is the fact that databases facilitate serendipity. Serendipity is often defined as a "happy accident" and it consists on discoveries that have a big part of chance in explaining them. Thanks to these databases scientists seem to be more likely to come with hazardous discoveries. Indeed, creativity, although depending on chance, can be induced by some favourable factors such an open mind of researchers or the accumulation of knowledge that allows scientists to identify what is odd and unexpected. In a similar logic, databases offer a complete view of the available data to the whole scientific community. This means that, on one side, scientists have a large number of elements to combine and a lucky discovery is more likely to happen. At the same time, the amount of scientists potentially doing this is big which increases even more the chances of an hazardous discovery.

5 CONCLUSION AND PERSPECTIVES

As we saw earlier, the use of large databases and the dependence of pharmaceutical industry on them is becoming more and more relevant in research. We have shown that this use does not only provide with a gain on productivity as one could thing, it also offers supporting conditions for creativity. Indeed databases put together knowledge coming from multiple disciplines and communities and make it accessible to everyone. They help to overcome problems related to knowledge distance. All of this is very relevant because databases are becoming crucial, not only for pharmaceutical research, but also for several other fields of science. The main contribution of this article has been to understand in which way the access to large public biological databases can favour scientific creativity. Databases can be, not only one more tool but also the resource that allows communities to share their knowledge and more likely be more creative. This is particularly relevant in a world where medical research is depending more and more on data. The use of large databases in medical research has a big potential in solving long term unsolved scientific challenges.

Another relevant aspect is the existence of favourable factors to serendipity. Serendipity has always been considered as a happy accident which allows for discovery. Although this "accident" was enabled by the fact that the person observing it had enough knowledge (or intuition) to consider it relevant; it was still considered a matter of pure luck. Our research suggests that databases could enable these kinds of accidents. The reason is that if we consider serendipity as the likeability to connect two pieces of knowledge which turn into a discovery, the more the pieces of knowledge an observer can understand, the larger the possibilities that she makes a connection. Also, the more observers analysing this pool of knowledge, the more the likeability that one of them makes a fructuous connection.

Another relevant contribution to theory is to show Knowledge codification as a possible source of creativity rather than the end of creativity. Codified knowledge is easier to understand and allows for the understanding of distant knowledge which is one of the biggest problems when it comes to knowledge management. There are however some drawbacks. One could think that because of the existence of standards and the presence of automation some interesting ideas could remain outside of the standards and produce an effect where some creativity is prevented because there is a lock in situation. Also, the existence of algorithms could generate a self-reinforcing effect where the same results (for instance combinations of pieces of knowledge) come out once and again and novelty is impeded. The access of everyone to all the available data could also create an effect where all researchers concentrate around the same topics, the low risk ones which allow for easy or fast results.

Further research should focus firstly on the long-term effects of Databases on creativity, for example the effect that databases have on the development of final products. It would be interesting as well to focus on the use of data at other stages of medical research such as clinical and pre-clinical trials.

6 REFERENCES

Arthur, W.B., 1989. Competing Technologies, Increasing Returns, and Lock-In by Historical Events. The Economic Journal 99, 116–131. https://doi.org/10.2307/2234208

Attwood, T.K., Gisel, A., Eriksson, N.-E., Bongcam-Rudloff, E., 2011. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European

Perspective, in: A. Mahdavi, M. (Ed.), Bioinformatics - Trends and Methodologies. InTech.

Audretsch, D.B., Feldman, M.P., 1996. R&D Spillovers and the Geography of Innovation and Production. American Economic Review 86, 630–640.

Bolser, D.M., Chibon, P.-Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V.D., Swan, D., Bassi, S., González, V., Suravajhala, P., Hwang, S., Romano, P., Edwards, R., Bishop, B., Eargle, J., Shtatland, T., Provart, N.J., Clements, D., Renfro, D.P., Bhak, D., Bhak, J., 2012. MetaBase--the wiki-database of biological databases. Nucleic Acids Res. 40,

Burt, R.S., 2004. Structural Holes and Good Ideas. American Journal of Sociology 110, 349–399.

David, P., 1985. Clio and the Economics of QWERTY. American Economic Review 75, 332–37.

Gong, S., Worth, C.L., Cheng, T.M.K., Blundell, T.L., 2011. Meet me halfway: when genomics meets structural bioinformatics. J Cardiovasc Transl Res 4, 281–303.

Heinze, T., Bauer, G., 2007. Characterizing creative scientists in nano-S&T: Productivity, multidisciplinarity, and network brokerage in a longitudinal perspective. Scientometrics 70, 811–830.

Heinze, T., Shapira, P., Rogers, J.D., Senker, J.M., 2009. Organizational and institutional influences on creativity in scientific research. Research Policy, Special Issue: Emerging Challenges for Science, Technology and Innovation Policy Research: A Reflexive Overview 38, 610–623.

Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery.

Hollingsworth, J.R., 2002. Research organizations and major discoveries in twentieth-century science: a case study of excellence in biomedical research.

Koskinen, K.U., Pihlanto, P., Vanharanta, H., 2003. Tacit knowledge acquisition and sharing in a project work context. International Journal of Project Management 21, 281–290.

Lee, Y.-N., Walsh, J.P., Wang, J., 2015. Creativity in scientific teams: Unpacking novelty and impact. Research Policy 44, 684–697.

Li, Q., Maggitti, P.G., Smith, K.G., Tesluk, P.E., Katila, R., 2013. Top Management Attention to Innovation: The Role of Search Selection and Intensity in New Product Introductions. Academy of Management Journal 56, 893–916. https://doi.org/10.5465/amj.2010.0844

Merton, R., Barber, B., 2004. The Travels and Adventures of Serendipity.

Munos, B., 2009. Lessons from 60 years of pharmaceutical innovation. Nature Reviews Drug Discovery 8, 959–968.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change. Cambridge MA and London: Harvard University Press.

Nooteboom, B., 2000. Learning by Interaction: Absorptive Capacity, Cognitive Distance and Governance. Journal of Management & Governance 4, 69–92.

Orsenigo, L., Malerba, F., 2015. The evolution of the pharmaceutical industry: Business History. Business History 57, 664–687.

Rodan, S., Galunic, C., 2004. More than Network Structure: How Knowledge Heterogeneity Influences Managerial Performance and Innovativeness. Strategic Management Journal 25, 541–562.

Schilling, M., Green, E., 2011. Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences. Research Policy 40, 1321–1331.

Smith, E.A., 2001. The role of tacit and explicit knowledge in the workplace. J of Knowledge

Stephan, P., 2012. How Economics Shapes Science — Paula Stephan | Harvard University Press

Takeuchi, H., Nonaka, I., 1995. The New Dynamism of the Knowledge-Creating Company 10.

Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical Combinations and Scientific Impact. Science 342, 468–472. https://doi.org/10.1126/science.1240474

Yaqub, O., 2016. Serendipity: Towards a Taxonomy and a Theory (SSRN Scholarly Paper No. ID 2841236). Social Science Research Network, Rochester, NY.

Zuckerman, H., 1967. Nobel Laureates in Science: Patterns of Productivity, Collaboration, and Authorship. American Sociological Review Vol. 32, No. 3 (Jun., 1967), pp. 391-403