Standardization and codification in the production of science: potential drivers of creativity

Nuria Moratal Ferrando
BETA. Université de Strasbourg

March 2016

Abstract

This paper attempts to show that data-bases collecting codified knowledge can lead to a wider and more varied knowledge base. For doing so, it explores the transformation of scientific results into data and the recent role of those large data-sets and related Science and Technology services on the production of science. The importance relies on the fact that the ability to use and combine more knowledge coming from different disciplines might be a key factor for creativity. I will as well show that the provision of those data bases and related services is meant to follow the rationale of Open Science and be freely accessible. For all of this we do a case study analysis. I use qualitative data from interviews combined with the analysis of policy documents, institutional reports and literature reading.

1 Introduction and theoretical background

Creation does not happen in isolation. The production of innovation requires recombining existing knowledge and ideas. Employees from one firm create and recombine ideas through a process of collective learning that happens within and across departments (Lorenz 1996: March 1991). But organizations do not only recombine internal knowledge. In their production process they use and combine knowledge that is coming from outside their boundaries (Rosenkopf and Almedia 2003). What all the fields of science and technology have in common when it comes to inputs needed for the production are cognitive inputs and external knowledge (Rosenberg 2004, Rosenberg and Nelson 1994).

These external sources of knowledge and the ability to exploit them is crucial for their competitive advantage (Dosi 1988: Singh and Agrawal, 2011). The idea of the criticality of knowledge flows, diffusion and recombination comes from the well known Marshallian externalities. But was some decades later when knowledge spillovers became the center of the Research Agenda (Grossman and Helpman, 1991; Lucas, 1988; Romer, 1990). The reasoning consists on the idea that research activities (public or private) spills over part of the knowledge produced. Universities have an important role here since they have always been the place for the creation of knowledge. And for the knowledge to be relevant it doesn't need only to be created but also transferred to the society. More precisely most schools of thought in economics agree that this knowledge has to be transferred to a specific sector of the society: industry.

However the simple fact of publishing knowledge doesn't make it accessible for the society. This knowledge has to be found (the agents need to know it exist) and understood. Furthemore, the ability to receive and profit from this knowledge spillovers is influenced by the knowledge distance (Audretsch and Feldman, 1996). What recently has been highlited by some schoolars is that excessively close actors might have little to exchange after a certain number of interactions (Boschma and Frenken, 2010). The production of new ideas needs from the combination of different pieces of knowledge that are related but at the same time complementary. However at some point agents risk to start combining and recombing the same kind of knowledge and it becomes redundant and les valuable and it lead to lock-in processes (Arthur 1989, David 1985).

I explore the impacts that the possibility of using a wider variety of knowledge has on science production. Codified knowledge is easier to understand specially when standardized, organized and available in archives or data-bases. I will try to show how this organization and codification makes knowledge easier to understand which allows scientist to use a wider variety of knowledge. This is the reason why I have chosen the case of bioinformatics. Bioinformatics codifies knowledge, establishes a common language and a way to express things that is common to people coming from different scientific disciplines. Then it makes this knowledge available in data-bases- This is why I look at bioinformatics as a tool to be able to understand a wider variety of knowledge which can be the first step towards more creativity.

To be able to understand why I am talking about the possibility to codify knowledge and store it in data-bases is important to keep in mind that in some scientific fields knowledge equals discovery. In the case I am studying here, which is biology, knowledge can be the discovery of the structure of a molecule. This can be the topic of a scientific paper or

a doctoral thesis. It is similar in other fields such us chemistry and physics.

2 Theoretical Framework and research gap

2.1 Absorptive capacity and production of knowledge

Creativity comes from the persistence on the combination (Simonton 2004)

Knowledge is a function of the knowledge base and the ability to create new knowledge by combing the existing one. The literature on absorptive capacity explains how not all the knowledge in the public domain is part of the knowledge base. Knowledge, even when published and available to everyone on the internet or public libraries is difficult to access. One needs to identify it and understand it. One needs to scan the environment and be able to find this knowledge. Just knowing this knowledge exist is already difficult. And even when found it has to be scanned, interpreted and learned (Cohen and Levinthal 1990). However that process can be improved with I&T services. A good example to better understand the improvement is thinking about the digital availability of journals. Before they were only available in paper and not all the universities had a big variety of journals in their libraries. Finding articles written about a specific topic was difficult and even when found there was a needed bureaucratic process to get a copy of it. Now some disciplines are experimenting something similar with the transformation of information contained in a scientific article into a piece of data.

In natural sciences, where knowledge creation consists in pure discovery the content of a paper can simply consist on expressing the structure of a specific protein or the observed interaction between that protein and a gen. This information previously explained in scientific papers is now available in a digital format. Digital knowledge is easier to find. Not only because you don't have to read a paper but also because the fact that it is in a database implies that it is classified and there is a search tool to easily identify specific gen or molecule and all what is known about it.

What is even more interesting here is that that codification crosses disciplines and scientific communities. For understanding each other people need to share certain basic perceptions. This requires a certain shared "interpretation system" (Weick 1984; Weick 1995) or "system of shared meanings" (Smircich, 1983), established by means of shared fundamental categories of perception, interpretation and evaluation inculcated by community culture. Even when looking at the same phenomenon different disciplines use different language, different interpretation systems and communication is complicated. With the codification of knowledge through data the barrier of language can be overcame.

I&T services play a role of codifying scientific results for an easy access and understanding, which directly affects the dimension of the knowledge base. The process needed to understand the relevant knowledge present in the public domain is not easy. The first step is to identify its existence. This very first step is not automatic, even when this knowledge is put in the public domain. Scientists need as well to be able to identify this knowledge as relevant and finally they have to be able to understand it, is what we call Absorptive Capacity (Cohen and Levinthal 1990). Absorptive capacity is important because no scientist or even group has enough cognitive resources to create knowledge. They

have to rely as well on others' cognitive resources and others' knowledge. The knowledge sharing happens within labs, communities and disciplines following this order. This has traditionally been the case in genetics where protein information was shared within networks (marx 2007).

The first aim of this paper is, therefore, to explore the recent and potential role of large knowledge data-sets and related Science and Technology services on the production process of science-intensive firms. In particular I want to show that these data bases and its related services can facilitate cross discipline combination.Data science and I&T services have allowed in the recent years a process of codification of scientific results by collecting them in data bases. These databases offer a large amount of detailed codified knowledge and are a way of technology transfer from the universities to the industrial sector. Improved data bases have allowed an ability to process larger amount of information to the point of being able to automatize part of the process during the production of knowledge. But is not only a matter of speeding up processes. These codification leads to a better cross-discipline combination of data which can lead, according to the literature on Absorptive capacity, to a boost in creativity. We will as well show how the provision of those data bases and related services is meant to follow the rationale of Open Science and be freely accessible.

2.2 On the role of Open Science

Since science is an accumulative process where big breakthrough discoveries are built up on smaller advances on the understanding of the world, an important input for the production of knowledge is previous knowledge or discovery. The system that makes previous discovery available for building up knowledge is the system of Open Science: disclosure of results, makes most of that information available in the public domain and usable by other scientists. The advances in I&T services over the past two decades have radically changed the way scientific results are inquired and accelerated research: internet libraries, digital repositories, etc. However nothing would have been possible without the social and political regime of Open Science. This regime has encouraged the dissemination of scientific results produced by government funded institutions. The way to do it has been by basically enforcing the traditional norms of science: reputation based on discovery and appropriation of the discovery based on priority of publication (Dasgupta and David 1994).

This is why I will try to show that, in the cases in which scientific knowledge can be translated into computational data and be stored and organized in data-bases, this databases should be freely available for everyone. This would allow for the spread of the knowledge and the possibility of more researchers (public or private) using this knowledge which was at the very beginning the goal of science funding by universities.

With the "Human Genome Project" this idea was applied at a large scale. The disclosure regime among participants was built upon the ideas of free and unrestricted access to each other's findings (Murray-Rust 2008). Even firms participated in the project and some related open data initiatives that emerged around this project (Pincock,2007; Thursby et al 2009; Allarakhia and Walsh, 2011). A generalization of these sharing practices instead of its limitation to certain projects would have numerous benefits to the science and therefore to the society.

2.3 Contribution

The new idea that this paper proposes and that is drawn through an interaction between the field study and the use of the current literature is the following. Codified and organized knowledge is easier to be understood even when knowledge distance is bigger. Therefore this codification and archive in data-bases can lead to a more varied (and also wider) knowledge base. This increase of variety in the knowledge base allows for the use of more varied cognitive resources when doing research which can lead to more creativity.

The first contribution of this paper to the existent literature in Economics of Science is relevant in different areas. In first place, the idea of codification of scientific knowledge, forgotten for some time, comes back to play an important role. I show how scientists facing the challenge of going through the inquiry the endless ocean of scientific knowledge (Fan et al 2014) can better lead with this problem thanks to data bases.

In a second stage this paper will try to show that these data bases should be publically and freely available. The most important reason is that there is a social interest on knowledge created by the society should be available to everyone that can create value for the society.

Summarizing I propose that the existence of big reliable databases and associated IT services allows for cross discipline combination. ALso that these data-bases and I&T services have to be freely available for an optimal provision and use.

3 Method and data

The research was undertaken utilizing a combination of document review and a case study consisting on qualitative data coming from interviews with key informants.Researchers have used the case study research method for many years in different disciplines but particularly in Social Sciences. Case studies consist on " analyses of persons, events, decisions, periods, projects, policies, institutions, or other systems that are studied holistically by one or more method. The case that is the subject of the inquiry will be an instance of a class of phenomena that provides an analytical frame — an object — within which the study is conducted and which the case illuminates and explicates." (Thomas 2001). Yin defines the case study research method as an "empirical inquiry for the investigation of contemporary phenomena in its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used" (Yin, 1984, p. 23).

Case study research is a research methodology first spread in the area of psychology but soon extended to other areas of social science such as management studies. Because of the nature and the tradition in these disciplines the approach is often inductive with the principle of building of theory from the data rather than using data to prove theory. In this paper the theory is built in a mixed way. The process begins being deductive but theory has been changed and enriched from data in a bidirectional process that goes from theory to data and from data to theory until there is something consistent built.

The reason why I chose this methodology is the nature of the subject. The focus on science-based industry makes it difficult to have quantitative information. Research is not the goal of these companies but the method to innovate. This research is therefore not systematically published and only a small part of it is reflected through patents. Moreover, qualitative methods have shown to be very good at getting rich information concerning processes that are too complicated to be shown in collectible data (Eisenhardt 1989, Yin 2011, Yin 2013) which I think is the case of this research. Case study research is useful and suitable when there is the need an understanding of a complex process. It allows for an emphasis is on the details and the contextual analysis. Its strengths relies on the possibility of looking at a variety of events or conditions and their relationships which is the only way to observe some phenomena that manifest in a variety of ways and can not be measured. Generality can not always be achieves and despite the bias that comes with the fact of being very sensible to interpretation is the only way to study some phenomena.

3.1 Introduction and perimeter of the case study

This study try to better understand some processes that happen within the science based industries. More specifically I use the case of the pharmaceutical industry.

Bioinformatics also has a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. This has the potential of increasing even more the level of codification because it has overcame big problems related to how different sub-disciplines use different names to relate to the same thing. This bring us again to the research question of this paper, codification offers access to a bigger amount and a wider range of scientific knowledge and this has the potential of increasing multi-disciplinarily and creativity.

Bioinformatics, as a whole, offers by itself the codification I have been talking about. However the discipline by itself is not enough and needs a large support system for the storage of data-bases containing this kind of information. There are a few databases in the world and I have decided to focus on only one of them because it makes it easier to contrast and to imagine contra factual situation to compare. The case chosen is EBI, which is considered by most of the studies and people consulted as the world leader on the provision of bioinformatics services.

3.2 History and origins of EBI

EBI stands for European Bioinformatics Institute and t is a world leader on the provision of this kind of services. Its origins lie in the first Nucleotide Sequence Data base that was stablished in 1980 at EMBL in Heildelberg, Germany. The initial goal was to stablish 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 directly receiving data. This required from highly skilled staff. In addition the magnitude of the database grew in scale when the Human Genome Project started. This gave it as well more visibility and therefore more use and more popularity. There was also a need for specific research activities. It is because all of this that the EMBL council decided, in

1992, to establish EBI which started working in 1996.

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 provides 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 large biological databases programmatically, eventually to build data analysis pipelines or to integrate public data with users' own applications. The 6 core data resources are operated by relatively large teams of 15 to 20 people (scientific curators, software engineers, bioinformaticians, and visitors including PhD students)[1].

3.3 Interviews

The main source of information consists of 19 interviews: 9 of them consist on exploratory large-scope interviews, mostly involving staff of EMBL-EBI. The aim was to understand the field of bioinformatics, how the databases are organised, how are they used, etc. 9 others were conducted involving heads of bioinformatics departments in pharmaceutical companies. Their objective was to assess the impact that the use of bio informatics has had on the way science is produced. Most of the interviews were conducted face to face and 2 of them were conducted via video conference. Finally there is an interview with a start-up and offering data services to the industry by using EBI resources.

All interviews were recorded and transcribed verbatim, under the conditions of anonymity and confidentiality of information. They lasted between 45 minutes and 2 hours. Anonymity conditions here implies not only not disclosing the name of the people and companies involved but also not disclosing any information that could lead to their identification.

TABLE 1. Case study interviews

Kind of interview	**Focus of interviews**
9 interviews with heads of bio informatics departments in pharmaceutical companies	Changes on the way they work, increase on amount and types of resources that can be used for their research.
9 interviews with EBI staff from the areas of: finance, external relations training and outreach	Large scope. Larger topics are covered
An interview with a start-up that offers data services to the industry by using EBI resources.	Large scope. Larger topics are covered

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

The first set of interviews were conducted during the years 2012 and 2013. It consisted on large-scope interviews that had the objective of getting to know the field of bio informatics and how the databases work. Impact of those databases is treated as well. Among the impacts the part we are interested on consists the one concerning the pharmaceutical industry. These interviews were used in different research project and not only for this research paper. Despite that fact very valuable information came from there since they provide a deep understanding of what their databases are, how are they managed and what are they exactly used for.

Table 1a. Content of general-scope interviews

Topics treated
Description and scope of the databases and the services offered by EBI
Description and scope of the databases and the services offered by EBI
How the databases are managed and human capital needs
Impacts on industry, university research, etc.

Concerning the industrial users of EBI that were interviewed, they were all first contacted by EMBL-EBI to get their agreement on being involved in the study, and then contacted by BETA. It consists on heads of Bioinformatics departments in large pharmaceutical industries. In particular they all work in companies that participate in a partnership with EBI that is called "Industry Programme". The Industry Programme of EBI is a kind of club that organizes, together with EBI, workshops and discussions about practical topics related to the use of EBI resources. This partnership is funded by the members via the payment of membership fees. The membership fee is, however, not the only cost of participating in this Partnership. Each member of the club sends 1 or 2 workers from the area of bio informatics to participate in 2 to 3 days meetings every quarter; as well as several training sessions during the year.

For anonymity reasons the name of the specific companies that agreed to participate in this study will be kept secret, as well as the name of the people that was interviewed. However what I can say about them is that they are directors of bio informatics Departments in the world's largest pharmaceutical industries. The name of all the companies that participate on this Industry Programme is public and easy to find in their website. The 9 interviewed for this study are among them.

The interviews were conducted between March and April 2015. The aim was to understand how processes have changed thanks to the use of EBI databases and contrast the previously exposed propositions of this paper. The topics treated start by the intensity on use of EBI resources. Since pharmaceutical companies have a long process of production I will look only at the relevance of these resources within the research area. This will take some questions because there is a lot of indirect use that is difficult to assess. In a second stage of the interview the questions will be aimed to know how relevant these resources are but not only in terms of direct or indirect use. Relevance translates in aspects like criticality and impact on final results of research. After that we approach the question

of how important bioinformatics is as a whole. The relevance of this lies on the fact that the discipline by itself implies codification and existence of data bases. However it is not an easy question to answer due to the difficulty to stablish a borderline between what is bio informatics and what is not. This is the reason why a case study is better than trying to assess bioinformatics as a whole. Finally we go to the important part that consists of which are the gains that codification offers to the process and which are the future perspectives and the potential of the discipline. The reason why all the interview is not focused only in this last part is because interviewees are reticent to give clear answers and often go for more vague ones.

TABLE 1b Contenct of specific interviews

Topics treated
Intensity and relevance of EBI resources for the Research performed in the company
Intensity and relevance of bioinformatics as a whole and. recent evolution.
Gains in terms of speed in processes and availability of resources. Future perspective and potential of bioinformatics

Finally, in December 2015 some extra interviews were conducted with all the 9 people already interviewed earlier in order to confirm their agreement with the information I had drown from the interviews and the conclusions I had reached.

3.4 Documents

Table 2. Reports, policy documents and desk research

Content
Final report of public consultation on Science 2.0 / open science, European Commission
Excellent Science in the Digital Age, European Comission
EVARio Reports available at http://evario.u-strasbg.fr/
American Economic Review, 2013. AER Data Availability Policy
Wellcome Trust, 2003. Sharing data from large-scale biological research projects:a system of tripartite responsibility. In: Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale,USA.

Documents used include institutional and European policy documents and reports, newspaper articles and academic journal articles. An important amount of general information came from EBI documents such as Annual Reports, Website, etc. Desk research on various Internet sources were extensively used as complementary sources of information. Finally the report of the EvaRIO research project was used as well. The project

focuses on impact of Research Infrastructure and it uses as well EBI as one of its case studies.

3.5 Literature review

n this section I will contrast the existing literature to the specificities of this case study and the topic treated. The first step is to look at the literature on public goods and check if the provision of databases and related services can be considered one. Here the literature is the widespread textbook literature on public goods that explains them through the characteristics of excludability and reality. In a second stage I will review some literature on public provision due to social welfare reasons and public interest. I will also look at some literature that treats the specific case of databases. This is however a difficult task because very little or nothing has been written in this area. This is why I will focus on the more general categories of public libraries and public archives. Finally there is a need for the review and comparison of all the literature on open access which is strongly linked to the idea of public goods.

4 Results

4.1 On codification and creativity

In this section I am going to talk about the findings of the interviews. Most of them come from the interviews with members of pharmaceutical companies. The rest of the interviews, where EBI staff were interviewed had as a main objective framing and understanding the mechanisms that make it possible to provide this kind of services. They also helped understanding the kind of data they provide and how these data are used across disciplines

The first part of the interviews consists on knowing how intensive the use of EBI resources is. The aim of the question is to verify that those resources are an important part of the production process. Otherwise it wouldn't be accurate to talk about crucial role of these resources. In first place the aim was to get a specific unity of measurement, consisting on hours of use. The idea was soon abandoned because of the difficulty interviewees had to answer it. Therefore the question is openly asked and the repeated scenario that can be observed in all the interviews is the following. It is hard to define a way to measure use of this resources because the internal systems of the companies have integrated data coming from EBI. All the companies studied have internal databases that are nourished with EBI resources.

In second place the aim is to find out how important EBI resources are for the Research department within the companies concerned. Again a repeated problem we have is that research, within the R & D is a really difficult category to define. Often there is not an R&D organization but a research and early development that runs from discovery of molecule all the way through to proof of concept in the patient, so phase 2. After that late stage development and that's a completely different organization. Exact figures are therefore a difficult thing to achieve. It has been possible, however, to get very useful

insight about the role of EBI data within this area.

Most companies have a very small number of people that do an intensive use of EBI resources. Here we are talking Bioinformatics is a general tool that everyone is using for a whole variety of things and in that sense all the companies said that they certainly provide tools for data interpretation and data analysis that could be called bioinformatics. However the use of databases and tools developed indoors using bioinformatics and using EBI services and data is widely spread within the research departments in pharmaceutical companies.

Concerning the access to EBI data one thing that is reputedly considered important is the education programme that the EBI has. Interviewees from small groups consider it helps them understand how the databases work because they do not have enough capacity to develop a wide range of in-house tools. That makes them less familiar with some databases and bio informatics tools and the training programs are very useful. There is a gap and those programs help raising the education level of the community and the establishment of standards and common language. This is a very important way of codification because it is a very diverse community with people coming from a variety of disciplines such as biology, pharmacology, chemistry, etc.

Traditionally people did not know how to program experiments or how to manipulate the data. What they did was to ask a statistician or a computer scientist for help. However the amount of data has grown so rapidly and its complexity as well that there is a need for the biologist to treat these data themselves. This has made the companies and the biologist invest in training and led to a situation where life scientists are able to use information resources to perform their research every day. The result has been a widespread use of these resources and the possibility of using very big amounts of data in their research. This way of working has become fundamental. Many interviewees talk about EBI being part of their fundaments.

The emphasis when talking about EBI resources is put therefore on the standards. These standards go from the way molecules or proteins are expressed, the way queries are computed and something as simple as the names. In all interviews this is repeated, the standards created have led to a very easy query of data and the rapid availability of data coming from very diverse sources that before couldn't be reached. For example, before EBI scientist called the same gens in different ways and it was very difficult to query data related to that specific gen. It was a lot of work to simply identify the labs that were working with a specific gen. Difficulty increased even more when the disciplines were different. Now the existence of EBI database has led to a standardization of names, even across disciplines, which makes it possible to scientists to access to a wider range of data.

Another key factor reputedly mentioned is the quality and the efficiency of the data. The fact that data are in one single source and they do not have to use time and resources on mixing different databases is important. But not only data, the data provided by EBI are high quality data since the providers and operators of the resources are as well users. They understand the data and therefore they can curate and clean them. Most interviewees agreed that many projects wouldn't be done if the data had to be collected, curated and standardized. Nearly all projects driven by bioinformatics means would not

have been possible without the public data available from the EBI. Without EBI they would use other less reliable sources, but without any public resource they could not work in bioinformatics or use bioinformatics during their research.

In summary, without the availability of this scientific data they would have to search through 40,000 journals to find the information required; this would just not be possible. More and more they are looking at the data more and not just the literature available as it is more informative, but it is only possible because of the large scale efforts to measure hundreds and hundreds and hundreds of data and putting them together. In the future, they think, that they will be doing the same with patient data.

4.2 On public provision and availability of resources

Goods are classified as public goods according to their characteristics and the two characteristics that define a good as public are non-rivalry and non-excludability. But there are not only public goods and private goods, there are a lot of groups in the middle that have only some characteristics of public goods. Club goods for example subtype of public goods that are excludable but non-rivalry, at least until reaching a point where congestion occurs. This means it is physically possible to exclude people from its use if they don't pay. However one additional person using this good doesn't imply higher costs. These goods are often provided by a natural monopoly. They are called as well natural monopoly.

Data bases, as archives or libraries are included in this kind of public good. Public goods don't necessary need to be publically provided and certainly the same happens with club goods. However they need regulation to avoid monopolies taking all the surplus and sometimes public provision of these goods is offered a solution for reaching a social optimal. This is specially recommended when there is a social or political interest (Coviello and Mariniello 2014)

The case of public libraries and similar archive resources have been shown as one in which, because its social interest, the solution to the public good problem should be the state provision of the good. This is due to the socially desirable effects that a higher use of the resources would offer (De Witte, Kristof and Geys 2011; del Barrio and Herrero, L.C. 2014). Innovation in the case study that concerns this paper is especially of social interest because it concerns discovery of new medicines and people's health.

4.3 Discussion and future research

One important implication for policy makers is the fact that, the financing of public research is as important as making this research available and understandable "for real" The knowledge pool that is in the public domain can indeed become codified and easily available. This would, of course, increase speed and productivity of the production of science. But is not only another improvement on speed thanks to better computing capacity. This can offer not only an automation of part of the process, it can boost innovation since it offers access to scientific knowledge that previously was not possible to access due to big gaps between disciplines. These traditional barrier to creativity can be overcome thanks to the codification and standardization of data and scientific results.

There are, however, several challenges to be faced. We are talking here about databases available regardless from which part of the world. Whose government should finance these centralized services is not an easy question to answer. Another important challenge consists on an ethical problem. Data sets and I&T services have both, public and private users. Public users will make their results public and collaborate to the enlargement of the databases. However private users will be able to save up big amounts of money and use these public services to help their own profit making.

There are as well some experiences of companies accepting to disclose data and basic research results, since this ones are still far from their marketable products. But it is a challenge since there will be always the risk and the feeling that they might be helping the competitor to develop some successful product. Existing research provides very little insight on how the firms could address this challenge. Most of the research about private-public partnerships look at specific agreements where there are normally IP issues related. There however some research on specific cases of data disclosure and some preliminary attempts trying to stablish a theoretical framework to study private participation on "Open Data" (Perkman and Schild 2015).

Immediate following research will consist on the development of a theoretical model that justifies the public provision of databases and after that the study of how to overcome the problem of the participation of private companies not only as consumers of the resources but as well as co-providers.

5 References

Lorenz, E. (1996) Collective learning processes and the regional labour market, unpublished research note, European Network on Networks, Collective Learning and RTD in Regionally-Clustered High-Technology SMEs

March, J.G. (1991) Organizational Consultants and Organizational Research. Journal of Applied Communication Research 19(1-2): 20-31.

Rosenkopf, L. and Almeida, P. (2003) Overcoming local search through alliances and mobility. Management Science 49(6): 751-766.

Dosi, G. (1988) Sources, procedures and microeconomic effects of innovation. Journal of Economic Literature 26(3): 1120-1171.

Singh, J. and Agrawal, A.K. (2011) Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires. Management Science 57(1): 129–150.

Romer P.M. (1990) Endogenous Technological Change. Journal of Political Economy98(5): 71-102.

Lucas, R. E. (1988) On the mechanics of economic development. Journal of Monetary Economics 22(1): 3-42

Grossman, G.M. and Helpman, E. (1991) Innovation and Growth in the Global Economy. Cambridge, Massachusetts: MIT Press.

Audretsch, D.B. and Feldman, M.P. (1996) R&D Spillovers and the geography of innovation and production. The American Economic Review 86(3): 630-640.

Cohen W.M. and Levinthal D.A. Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly, 35(1):128–152, March 1990

Arthur, W.B. (1989) Competing Technologies, Increasing Returns, and Lock-in by Historical Events. Economic Journal 99: 116-131.

David, P.A. (1985) Clio and the Economics of QWERTY. American Economic Review 75(2): 332-337

Weick Karl E. Toward a Model of Organizations as Interpretation Systems. The Academy of Management Review. Vol. 9, No. 2 (Apr., 1984), pp. 284-295

Keick Karl E. The Mann Gulch Disaster: the Collapse of Sense-Making in Organizations. Administrative Science Quarterly, 1993

Smircich L. Concepts of Culture and Organizational Analysis. Administrative Science Quarterly. Vol. 28, No. 3, Organizational Culture (Sep., 1983), pp. 339-358

Cohen and Levinthal. Absorptive Capacity: A New Perspectiveon Learning and Innovation. Administrative Science Quarterly, 35(1):128–152, March 1990

Marx. J. Trafficking Protein Suspected in Alzheimer's Disease. 315(5810):314–314, January 2007.

David Paul.A. and Dasgupta P. Toward A New Economics of Science. Research Policy, 23(5):487–521, 1994.

Murray-Rust . Open Data in Science : Nature Precedings 2008

Pincock. A Role for Mathematics in the Physical Sciences. Noûs, 41(2):253–275, June 2007.

Thursby, J; Fuller Anne W; and Thursby, Marie . US faculty patenting: Inside and outside the university. Research Policy, 38(1):14–25, 2009.

Allarakhia Minna and Walsh Steven. Managing knowledge assets under conditions of radical change: The case of the pharmaceutical industry. Technovation, 31(2–3):105–117, February 2011.

Fan, J; Han F, Liu H. Challenges of big data analysis. National science review, 2014

Kathleen M. Eisenhardt and Melissa E. Graebner. Theory Building From Cases: Opportunities And Challenges. Academy of Management Journal, 50(1):25–32, February

2007.

Robert K. Yin. Applications of Case Study Research. SAGE, June 2011.

Robert K. Yin. Case Study Research: Design and Methods. SAGE Publications, Inc, Los Angeles, fifth edition edition edition, May 2013.