Paper to be presented at the DRUID Academy Conference 2016 in Bordeaux, France on January 13-15, 2016

Big Data and Codification in knowledge intensive industries. A case study on the Pharmaceutical Industry

Nuria Moratal
University of Strasbourg
BETA
nuria.moratal@gmail.com

Abstract

[[State of the art]] The aim of this paper is to explore role of codification of knowledge on the production process of knowledge-intensive industries. Advances in computing and storing capacity have allowed in the recent years for a process of codification of scientific results by translating them into computational data and collecting them into databases. These databases offer a large amount of detailed pieces of knowledge and the possibility of computing them together and combine them in many ways. [[Research gap]] Literature on Big Data and I&T services explains how not only we have now the possibility of computing larger amounts of data, Big Data also allows for the combination of multiple information that we didn't use to consider as data, as well as the combination of more varied data that we didn't use to compute together (Inmon and Linstedt 2015). Despite the big boost on research concerning Big Data little attention has been payed to the specific case of Big Data in the production process of knowledge and therefore the impact that these advances in computation have on innovative industries. Following the literature on Absorptive Capacity, I argue that the codification that comes with Big Data can allow for a combination of more diverse knowledge and therefore it can enable creativity and be the first step towards more innovative industries. In a second stage of the article I defend as well the need for these databases to be freely available and offered by non-by-profit driven institutions. [[Theoretical arguments]] Knowledge production is a function of the knowledge base and the ability to create new knowledge by combining all the previously acquired one. The knowledge base consists on all the knowledge that the agent has and it depends on what is known as the Absorptive Capacity. Absorptive Capacity is the ability of the agent (for instance a firm) to identify relevant knowledge in the public domain and understand it. This process is difficult and requires of scanning, interpreting and assimilating (Cohen and Levinthal 1990). The more similar the knowledge is, the more chances there are to find it and understand it. On the other hand, the more different this knowledge is, the more possibilities of being creative and finding something new (Howell 1999; Patel and Pavitt 1994; Hervas-Oliver et al 2012). There is a need for balance. The codification that comes with Big Data makes distant knowledge more understandable. This, together with the possibility of using I&T tools to process and combine it can shift the balance towards a higher ability to use varied knowledge and therefore an increase on creativity during the knowledge production process. This doesn't invalidate previous research showing that a firm needs to do research itself to be able to understand the knowledge produced by universities and use it in her production process. It means only that the diversity of knowledge that they can

learn is higher and when looking for new knowledge in the environment firms can aim for more distant knowledge. [[Method and data]] The methodology used is qualitative key study methodology. The case studied is the one of the Pharmaceutical Industry using Bioinformatics databases provided by EBI (European Bioinformatics Institute). The research was undertaken utilizing a combination of desk research and interviews with key informants. The data used consists of 20 interviews: 10 were conducted involving heads of bioinformatics departments in the world leading pharmaceutical companies. Other 10 interviews involve staff of EBI, in charge of managing the data provided. Regarding the documents used they include institutional and European policy documents and reports, newspaper articles and academic journal articles. [[Results]] I have observed a shift of paradigm: agents now do use more varied knowledge than before. Not only they are able to use knowledge produced by other disciplines than their own, they are also able to use knowledge produced by different communities within the same discipline. The main gain comes from having reached big levels of standardization in nomenclature of genes, proteins and molecules. This standardization makes possible to identify knowledge that previously was not identifiable. Since these databases were created researchers from different communities speak now the same language and that allows them to use each others resources. A good picture of how important this situation is, is the comparison with other scientific fields (mainly chemistry) were the absence of such a level of codification makes cross community combination impossible. In addition I have found some insights in the existing literature showing that these databases containing scientific knowledge should be provided publicly because their characteristics show that is the only way to reach a socially optimal use. I use for that existing models for goods with similar characteristics but I intend, in a future paper, to built a specific model for the case of databases containing scientific results. [[References]] 1- W. H. Inmon and Daniel Linstedt. 2.1 - A Brief History of Big Data. In W. H. InmonDaniel Linstedt, editor, Data Architecture: a Primer for the Data Scientist. Morgan Kaufmann, Boston, 2015. 2- Wesley M. Cohen and Daniel A. Levinthal. Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly, 35(1):128–152, March 1990. 3- Jeremy Howells. 5 Regional systems of innovation. Innovation policy in a global economy, page 67, 1999. 4- Parimal Patel and Keith Pavitt. The continuing, widespread (and neglected) importance of improvements in mechanical technologies. Research policy, 23(5):533–545, 1994. 5- Jose-Luis Hervas-Oliver, Jose Albors-Garrigos, and Juan-Jose Baixauli. Beyond R&D Activities: The Determinants of Firms' Absorptive Capacity Explaining the Access to Scientific Institutes in Low-Medium-Tech Contexts. Economics of Innovation and New Technology, 21(1-2):55–81, January 2012.

Jelcodes:O32,H40

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

Nuria Moratal Ferrando
BETA. Université de Strasbourg

December 2015

Abstract

This paper attempts to show that codification of scientific 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

In Economics is widely accepted that growth depends on the creation of knowledge(OECD 2002). We consider the production of knowledge as a function of the knowledge base and the ability to create new knowledge by combining the existent one. The knowledge base consists in all the knowledge that an individual or a group have. And it is by combining their existing knowledge in several ways that they can discover new things and reach new knowledge

It is through this knowledge base that societies will find new efficient ways of working. Universities have always been the place for the creation of knowledge. However 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 : the industry. This is possible thanks to a regime of open science and appropriation of the knowledge production via publication.

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. We are only able to understand knowledge that is not too distant from our own knowledge. Knowledge distance is, in fact, a key factor. We tend to use knowledge that is not very different to our own knowledge because we can understand it better. However this limits creativity. In order to be creative we need to use varied knowledge.

I explore the possibility of using a wider variety of knowledge thanks to its codification. Codified knowledge is easier to understand. That is why I have chosen the case of bio informatics. Bio informatics codifies knowledge, establishes a common language and a way to express things that is common to people coming from different scientific disciplines. 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.

The aim of this paper is, therefore, to explore the recent and potential role of large data-sets and related Science and Technology services on the production process of science-intensive firms. In particular I want to prove 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 Theoretical background and research gap

2.1 The production of Knowledge and the Absorptive Capacity

In economics we consider science production as 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. It takes time and it is not obvious (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. Well now we 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 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).

2.2 The Open Science institution and the ideal of open data

Factual data are a major resource for the production of science and therefore fundamental for our innovation systems. Research laboratories and science oriented firms not only rely on a collection of in-house observations. Since science is an accumulative process where big breakthrough discoveries are built up on smaller advances on the understanding of the world, an important part data consists on previous scientific discovery. The system of Open Science: disclosure of results, makes most of those data 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). Recently unprecedented opportunities have been created and the increase in speed and stocking capacity has open potential ways for an even faster inquiry of scientific results. Now any kind of scientific result can be translated to electronic data and stored in electronic database. These data can be computed in order to create a sort of digital laboratory. In other words, we have the possibility to automatize an important part of the knowledge production process. We could talk about an enlargement of the knowledge base. The result has been the acceleration of some processes and the possibility of more complex studies (Mount and Padney 2005).

The idea of Open Data as a rational evolution of the idea of Open Science promises a future were knowledge production is facilitated even more. If after doing the effort of collecting the data those are made public the investment of the society will have higher returns because more people will be able to use them in a creative way. However, even the Open Science regime has led to a situation in which researchers are encourage to keep their data collections secret. There are several reasons for this going from the fear to have their studies questioned to other scientists doing the research they planned to do. But despite this traditional of opening up scientific results but keeping data secret a lot of initiatives for Open Data have rose.

The idea of Open Data in science has been discussed since long time ago in leading journals of diverse scientific fields. However it is with the “Human Genome Project” that is 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. It allows for the replication of experiments, the correction of errors and expand existing research. Furthermore it makes possible to compile data and codes making it easily findable, available and understandable to other researchers (Andreoli-Versbach and Mueller-Langer 2014). Not only it will boost research, it would as well increase innovation and improve the whole global science system.

All these benefits that a widespread availability of data would offer are related to the way science and knowledge products are produced. The production of knowledge requires having a big amount of previously acquired knowledge and an ability to understand the knowledge in the public domain. Innovation requires creativity which is the ability to combine the two of them.

2.3 Theoretical arguments

he 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. In second place we show how scientists are facing the challenge of going through the inquiry the endless ocean of scientific knowledge (Fan et al 2014). I propose that (i) The exitence of big reliable databases and associated IT services allows for cross discipline combination (ii) These data-bases and I&T services have to be freely available for an optimal provision and use.

3 Methodology

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. Now it is the more and more common to find it in articles concerning the Economics of Science. Qualitative methods have shown to be very good at getting rich information concerning processes that are too complicated to be shown in collectible data (Einsenhardt 1989, Yin 2011, Yin 2013). 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.

Is the case for the phenomenon that we are interested on here. In principle it seems quite easy to get systematic quantitative information on whether or not knowledge creating agents are using a wider variety of data or not thanks to databases. However after starting this research I realized it is not that simple. In first place production process of knowledge is long and the inputs required are varied. Drawing the lines between what was data, and what wasn't was not an easy task. In second place as economist we need an interaction with an interviewee to understand how the process of producing knowledge in that specific field works.

.

3.1 Introduction and perimeter of the case study

This study will look at the science based industries, more specifically it focuses on the pharmaceutical industry and the Bioinformatics, which is a multidisciplinary field that combines disciplines such as computer science, statistics and mathematics to understand biological data. The discipline uses programming mainly for the study of genetics and related fields. This is the reason why the study is adequate for the research question approached by this paper. The use of programming methods implies the use of a highly codified language. Few scientific disciplines and its related methodologies have reached the level of codification that we see in the case of life sciences and biology. It is, as well, a data intense discipline.

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 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. Other 9 interviews are used as well. It consists on large-scope interviews, mostly involving staff of EMBL-EBI: people responsible for external relations, administration and finance, outreach and training. 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

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

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

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 1a 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

The second 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 1b. 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.

3.4 Documents

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.

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.

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.

References

[1] Evaluating the efficiency of museums using multiple outputs: evidence from a regional system of museums in Spain - International Journal of Cultural Policy - Volume 20, Issue 2.

[2] Pharma goes open access | The Scientist Magazine®.

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

[4] Wesley M. Cohen and Daniel A. Levinthal. Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly, 35(1):128–152, March 1990.

[5] Decio Coviello and Mario Mariniello. Publicity requirements in public procurement: Evidence from a regression discontinuity design. Journal of Public Economics, 109:76–100, January 2014.

[6] Richard L. Daft and Karl E. Weick. Toward a Model of Organizations as Interpretation Systems. The Academy of Management Review, 9(2):284–295, April 1984.

[7] Jianqing Fan, Fang Han, and Han Liu. Challenges of Big Data Analysis. National Science Review, 1(2):293–314, June 2014. arXiv: 1308.1479.

[8] Anil K. Gupta and Vijay Govindarajan. Knowledge flows within multinational corporations. Strategic Management Journal, 21(4):473–496, April 2000.

[9] David W. Mount and Ritu Pandey. Using bioinformatics and genome analysis for new therapeutic interventions. Molecular Cancer Therapeutics, 4(10):1636–1643, October 2005.

[10] Peter Murray-Rust. Open Data in Science. Serials Review, 34(1):52–64, March 2008.

[11] Yiming Qin, Hari Krishna Yalamanchili, Jing Qin, Bin Yan, and Junwen Wang. The Current Status and Challenges in Computational Analysis of Genomic Big Data. Big Data Research, 2(1):12–18, March 2015.

[12] Kristof De Witte and Benny Geys. Evaluating efficient public good provision: Theory and evidence from a generalised conditional efficiency model for public libraries. Journal of Urban Economics, 69(3):319–327, May 2011.

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

[14] Robert K. Yin. Case Study Research: Design and Methods: Design and Methods. SAGE Publications, May 2013.