International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 7 No. 4 Aug. 2014, pp. 1251-1273

http://www.ijias.issr-journals.org/

"Bigger Data" Visualization to Visual Analytics: a path to Innovation. "Happening, definitely! Misleading, possibly?" A review of some examples applicable to IP Discovery

Damien Lapray[1] and Serge Rebouillat[2]

[1]Currently at EPFL, CH1015 Lausanne, Switzerland

[2]Currently at DuPont Int., CH1218 Geneva, Switzerland

ABSTRACT: An image is worth a thousand words. This is a common adage which may have to be revisited. The query "Eiffel Tower" yields about 60 million images using Google™ search engine. These words combined with "steel structure" score about 20 000 images.

The power of images is paramount. With about 80 million enforceable patent documents, a large number containing images, one may wonder whether the adapted tools to exploit this image databank are available and used.

Adding three dimensional activation of patent drawings by means of computer aided design would likely return creative amazements with large potential for innovation. In 2010 ideators filed approximately 2 million new patent applications around the world. These patents tend to contain more readily exploitable images.

Combinatorial, associative or intersecting approaches, as illustrated in the introduction, are definitely a major source of inspiration for innovation, moreover disruptive.

What about the "Big Data" necessity? Can the 60-70's technology wonders, such as PCs, Biotech, Mechatronics, further evolve today without the "Big Data" component?

Big data approach is definitely not common in the IP domain; matter of legal fears or lack of adapted tools? The question will anyway probably not slow-up the advent of Big Data in a broad fashion in many areas.

Inclusive innovation, with a goal to serve beyond the development mainstream, encompasses more consumer data therefore Big Data analysis too. Inclusive, open and disruptive innovation modes are pending on good and clear visualization of the trends, initially partly or mostly technological.

This chapter, as part of a series on innovation, attempts to answer some questions related to the above matter and provides insight in the visualization technical status and its potential and direct applicability to IP analysis, and IP discovery in general. Visual analytics, although not developed, are integrated in the horizon of a bigger data analysis bringing additional questions such as:

Beyond the classical synergy -additive- equation, is there a potential for multiplying the ideation outputs?

Furthermore, is there presently too much emphasis engaged on the data itself, rather than the analytical trends and the acumens that can be produced? Are the available tools, such as for extraction, suitable?

KEYWORDS: Innovation - open, disruptive, inclusive -, visualization, Big Data, visual analytics, ideation, trend analysis, neural networks, collaborative, Collaboratory™, adjacent technology analysis, ATA©, IP strategy, semantic analysis, image analysis, artificial intelligence, reverse engineering, confidence, curiosity, combinatorial, biotechnology, energy.

Disclaimer: This article is primarily for educational purposes. Selected cases are strictly illustrative. Neither the author nor the illustrator assumes any liability for any errors or oversights, or for how this article or its contents are utilized or interpreted, or for any consequences resulting directly or indirectly from the usage of it.

This assessment is intended to educate and raise awareness of some of the complex issues that surround the intellectual property in the field of knowledge extraction from the about 80 million patent documents available, and to assist in the development of practical skills for

Corresponding Author: Serge Rebouillat

dealing with inventions. It does not seek to provide legal, managerial or technical advice on intellectual property related law as such. For any guidance, legal or any other, seek advice from the appropriate professionals; this study can by no mean substitute for expert legal, technical and managerial advice.
The opinions expressed by the writers in this article do not necessarily represent the viewpoints of the companies the author are employed by.

1 INTRODUCTION

An image is worth a thousand words, is a common adage which may have to be revisited. The words "Eiffel Tower" yield about 60 million images using Google™ Images search engine. These words combined with "steel structure" scores about 20 000 images.

The power of images in our day-to-day activities as well in professional situations, being scientists, engineers, economists, business strategists is dominant. With about 80 million enforceable patent documents, a large number containing images, one may wonder how the professionals cited above cope without adapted tools to exploit this image databank. Adding three dimensional activation of patent drawings by means of computer aided design would likely return creative amazements with large potential for innovation. The growth of patent applications is substantial; in 2010 ideators filed approximately 2 million new patent applications around the world. These patents tend to contain more readily exploitable images.

Combinatorial, associative or intersecting approaches, as illustrated on Fig. 1, both self-explanatory, are definitely a major source of inspiration for innovation, moreover disruptive. Adding questioning, observation, networking and experimenting will complete the stereotypical disruptive innovator profile, as reproduced on Fig. 2 [1] (Dyer et al., 2009). The adjacent technology analysis serving in this particular situation as an evergreen soundboard and a discovery lens, ATA© has been rather abundantly covered by Rebouillat.

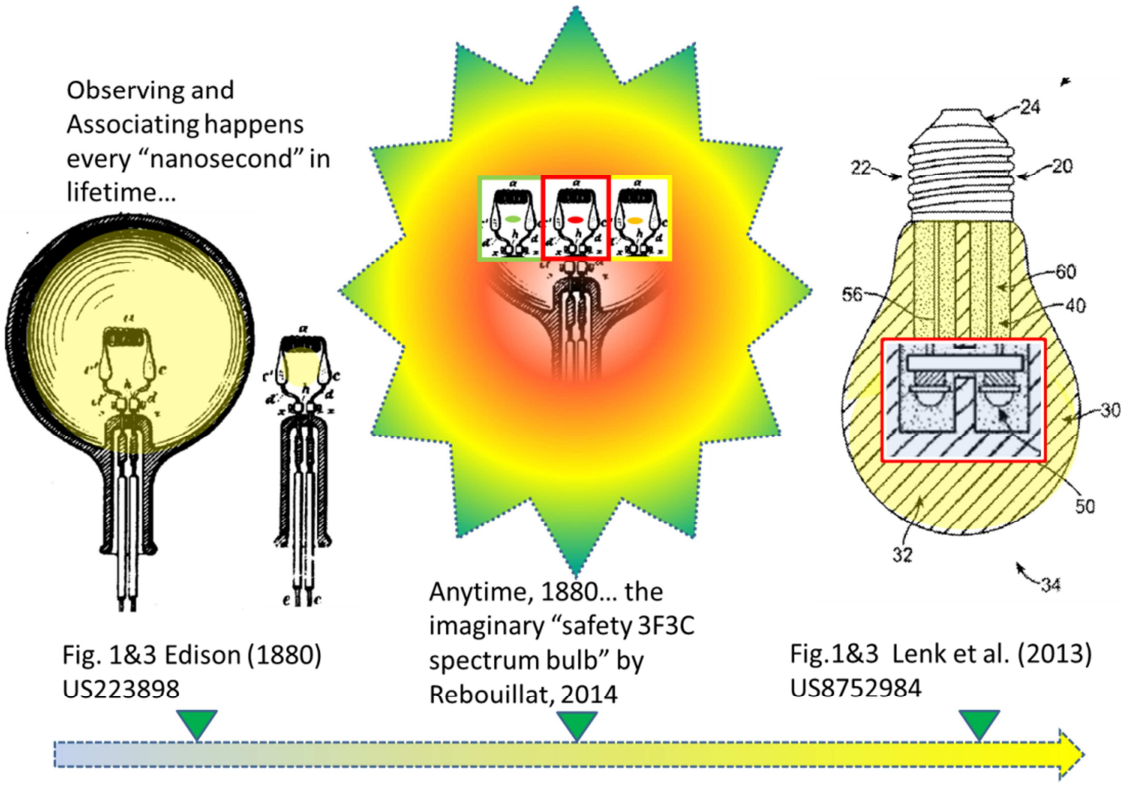

Fig. 1 Combinatorial, associative or intersecting approaches, as illustrated above are part of the Disruptive Innovation Path: from monofilament bulb to multiLED light emitting systems, there are multiple inspiring options simply derived from observation and visualization, i.e. spontaneous "visual analytics".

Also:

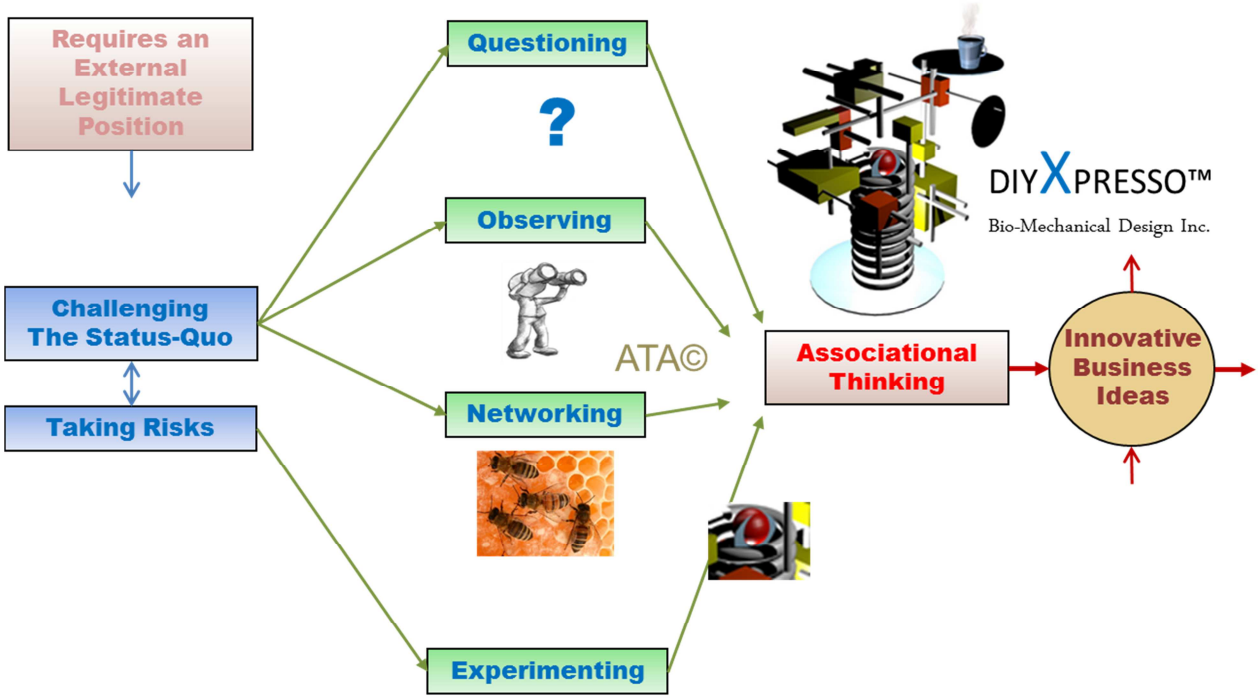

Fig. 2 Adding questioning, observing, networking and experimenting can complete the stereotypical disruptive innovator profile. "Big Data" is revolutionizing the adopted skill requirements.

What about the "Big Data" necessity? Will the disruptive innovator criteria, based on the 60-70's technology wonders, such as PCs, Biotech, Mechatronics[1] apply today without the "Big Data" components? Is this question motivated by provocation or does it spring from farsighted logic?

Big data approach is definitely not common in the IP domain; matter of legal fears or lack of adapted tools? The question will anyway probably not slow-up the advent of Big Data in a broad fashion in many areas.

Inclusive innovation, with a goal to serve beyond the development mainstream, encompasses more social consumer data therefore "Big Data" analysis too. Inclusive, open and disruptive innovation modes are pending on good and clear visualization of the trends, initially partly or mostly technological.

This chapter, as part of a series on innovation, attempts to answer some questions related to the above matter and provides insight in the visualization technical status and its potential and direct applicability to IP analysis and IP discovery in general.

Is there a potential for multiplying the ideation outputs beyond the classical synergy additive equation?

Is there presently too much emphasis engaged on the data itself, rather than the analytical trends and the acumens that can be produced? Are the available tools for extraction suitable? Can Visual Analytics, "the science of analytical reasoning facilitated by visual interactive interfaces", help and when? These are additional questions.

Furthermore and additively:

Part of our economic model relies on innovation processes (Fig. 3) to fuel industrial growth.

[1] *Trademark applied by a company in Japan with the registration number of "46-32714" in 1971.*

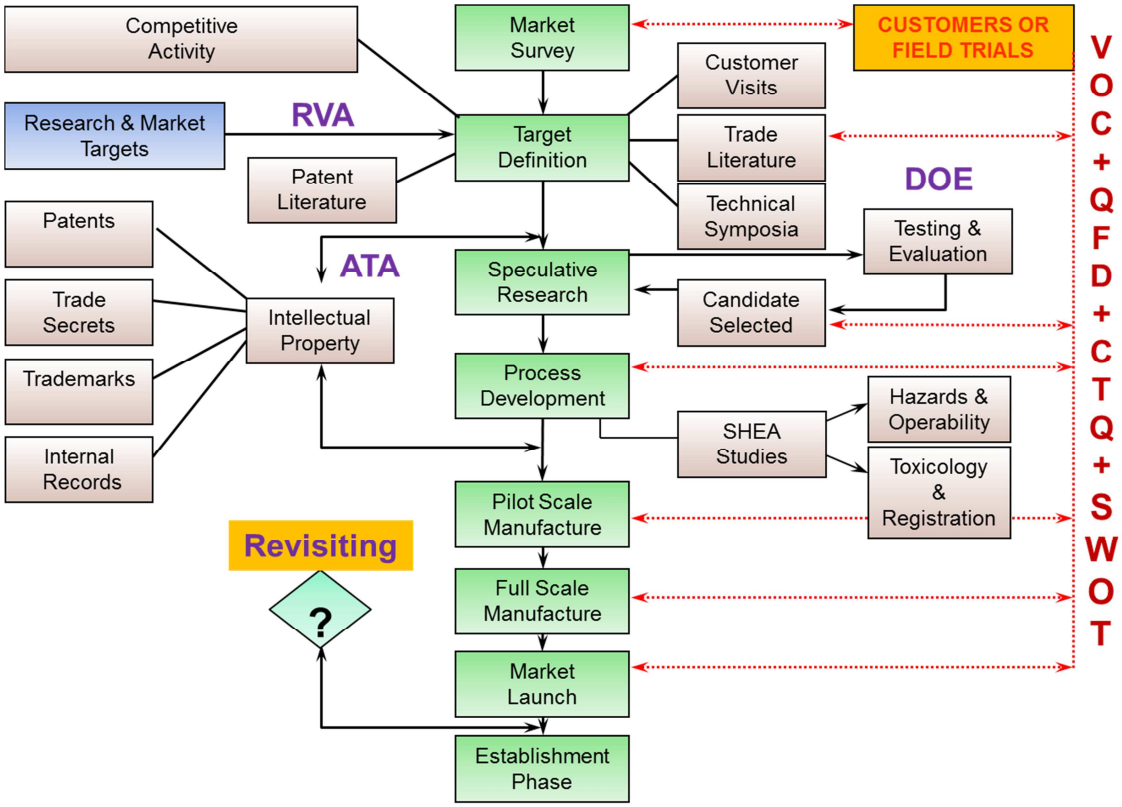

Fig. 3 The Z-Process, above, combines multiple business and technology management processes.

Until recently innovation was restricted to the development of ideas and concepts within firms' environment. However, this model has been challenged during the past decade with the arrival of "open innovation", a "paradigm" suggesting that "firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology" [2] (West & Bogers, 2011). This paradigm generated multiple business models [3] (Rebouillat, 2013) and ideas to make the best of the available material. In parallel to this conceptual revolution, the amount of the data one can access and use to stimulate his creative imagination grew exponentially. Innovation, once restricted to a few creative minds is now an open market. In the intellectual property (IP) domain, the number of granted patents is growing every day following a slope never encountered before. From 2008 to 2010 the World Intellectual Property Organization (WIPO) received about 5,000 new patent applications per day, and design patents in China were more than 500,000 in 2012. It is accompanied by an explosion in the number of scientific productions. In 2010 Medline received about 110 new articles per day. Furthermore, areas such as biology collect and share petabytes of data through public repositories such as the European Bioinformatics Institute or the US National Center for Biotechnology Information [4] (Marx, 2013). This explosion of data leaded to the creation of the "Big Data" concept and its associated real-time decision-making tools. However, Big Data *per se* yield neither meaning nor value. Many innovative opportunities are wasted because we lack the ability to extract and process this large amount of data. New promising solutions are emerging to tackle the associated challenges. One of the most promising domains of research is "visual analytics" where analytical reasoning is facilitated by interactive visual interfaces. This whole new scientific domain promotes collaborative work and human-machine interaction to create the foundation of new innovative paths.

The aim of this paper is also to continue recent works [3], [5], [6] (Rebouillat & M. Lapray, 2014; Rebouillat & D. Lapray, 2014; Rebouillat, 2013) and go further in the understanding of the use of visualization-based data discovery tools in the IP domain.

2 DATA ANALYSIS AND INNOVATION

2.1 OPEN INNOVATION

In the past, companies' strategy would be to accumulate IP to provide design freedom to their employees and to avoid costly litigation.

More precisely and more recently revised to:

1- Maintain Superior Competitive Position
- Minimizing elapsed time from Idea to market
2- Ensure Business Flexibility
- Maintaining and expending the freedom to operate
3- Secure Business Profitability
- Minimizing legal/competition exposure

However, some if not most patents were worth very little and as a result were never used by the business that held them. Ever faster technology cycles and stronger global competition prompted Henry Chesbrough to propose a new paradigm, so called "open innovation" [7] (Chesbrough, 2006). It promotes the use of external ideas in addition to internal ones in order to advance technological growth. In Open Innovation, "IP represents a new class of assets that can deliver additional revenues to the current business model, and also point the way towards entry into new businesses and new business models" [7] (Chesbrough, 2006).

For patent strategy, IP awareness and vigilance is critical: offensive and/or defensive, creating value and/or distributing value. The challenges between creating value opportunities and strictly controlling and distributing value, subtend the business opportunity within its rather fast evolving boundaries [3] (Rebouillat, 2013).

In order to detect the broadest range of possible options for introducing new knowledge/expertise into an organization, it is necessary to process a large amount of data from many different sources. Using data mining technologies the professional has at hand a multitude of possibilities to generate useful information. A huge volume of diversified data is now available and associated tools are being developed to extract precious information [6] (Rebouillat & D. Lapray 2014).

2.2 BIG DATA

The definition of Big Data does not only stand for huge Volume; the novelty also comes from its Velocity and Variety. The "three Vs" of Big Data represent the challenges ahead when one wants to make sense of the available information [8] (Intel, 2013). It is not the aim of this paper to discuss the new questions associated to the exploitation of Big Data such as the dangers for privacy, a putative obsolescence of the scientific method [9] (Anderson, 2008), an evolution of knowledge, outsourcing to distant service provider, etc. The reader can find an excellent summary of all these ideas in a report of the Aspen Institute [10] (Bollier, 2010).

In the context of Big Data the consensus among all professionals is that it can help identify emerging trends, improve business decision making and develop new revenue-making strategies. There are countless opportunities to make scientific research more productive, and to accelerate the process of discovery and innovation. The latter being the key matter of this paper. How can businesses use this novel material to its best in order to boost creativity?

The amount of data to analyze is so immense that traditional tables and charts cannot help the researcher to find relevant patterns of information. The whole field of knowledge extraction and management is in need of new solutions. Mapping technologies seem to carry a lot of potential where the researcher uses different visual formats and its knowledge to detect and interpret interesting correlations. Bill Stensrud, Chairman and CEO of InstantEncore, said, "I believe in the future, the big opportunity is going to be non-human directed efforts to search Big Data, to find what questions can be asked of the data that we haven't even known to ask" [10] (Bollier, 2010). There are many ways to take advantage of the variety of data especially for IP researchers [6] (Rebouillat & D. Lapray 2014). We believe that visual analytics is one area that holds a lot of promises for the IP domain and will be discussed further in the rest of this paper.

3 VISUAL ANALYTICS

3.1 A "HUMAN FRIENDLY" DATA REPRESENTATION

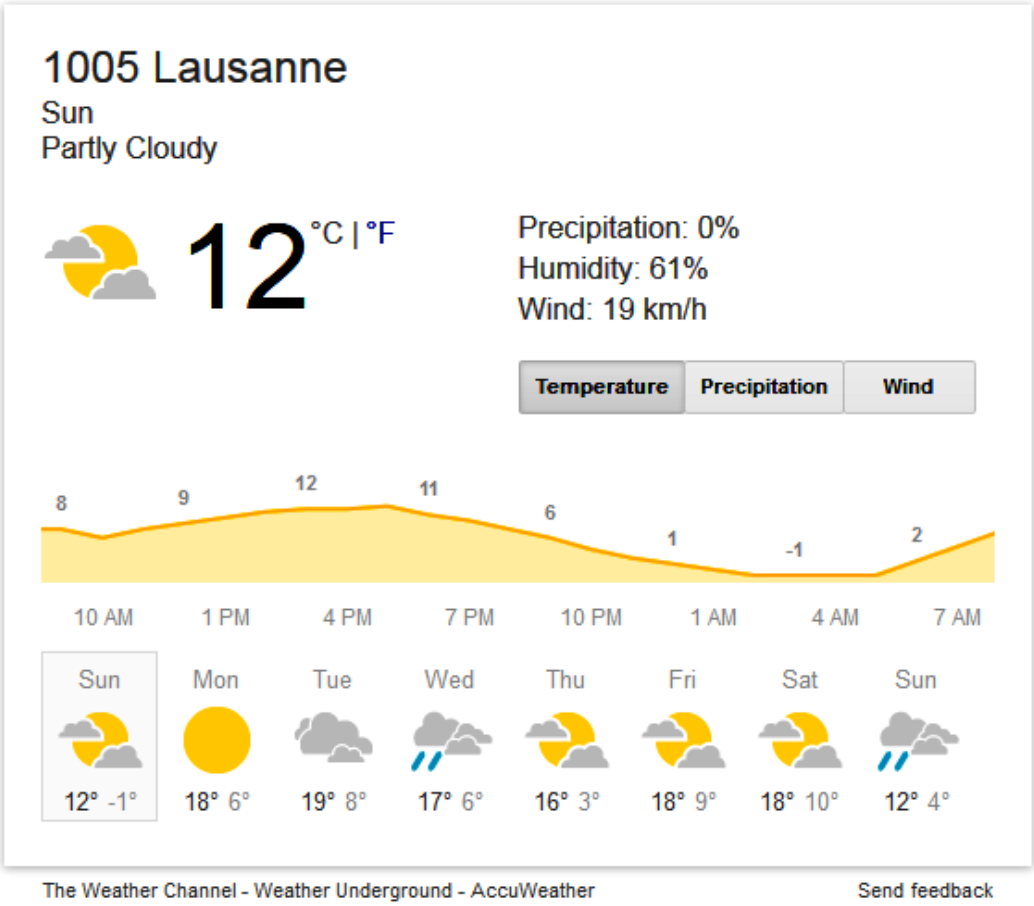

Fig. 4 Visual analytics helps in daily life activities where list of data would not allow grasping relevant information quickly. Here the weather forecast as displayed by Google™ search engine.

Since the birth of digital data we have been used to manipulate data in a non-intuitive way for humans. List of numbers or characters is perfectly understandable by a computer however it is another story for our brain. That situation was manageable when data where sparse and easy to manipulate. Is it still possible considering the volume of information? The researcher cannot easily find relevant information in complex data tables and many discoveries may remain unnoticed. It is becoming crucial to present data in a more "human friendly" way. Visual analytics is changing the way we interact with data. This tool is not new and has been around for a while. Weather forecast data are a perfect example of its use (Fig. 4). It is more than a simple esthetical tool, it's strength lays in the way this technology combines human and electronic data processing abilities [11] (Keim et al., 2010). Human-machine full cooperation is the key to innovation [6] (Rebouillat & D. Lapray 2014). Visual analytics is a scientific field on its own and this is out of this paper's scope to cover all aspects of it (for more information see [11] Keim et al., 2010). The two main areas that will be covered concern the use of visual analytics in the exploration and presentation of data. Both are highly intermingled and used through the whole process. They will be differentiated for clarity reasons. The use of visual representation of data is a fantastic opportunity to promote collaborative work (Fig. 5). It is indeed easy to sit around a map and think of the next step, a mission far more tedious with a table containing thousands of entries. The relative simplicity of these tools and the intuitive way people can navigate through millions of data points make this technology extremely attractive in the innovation process.

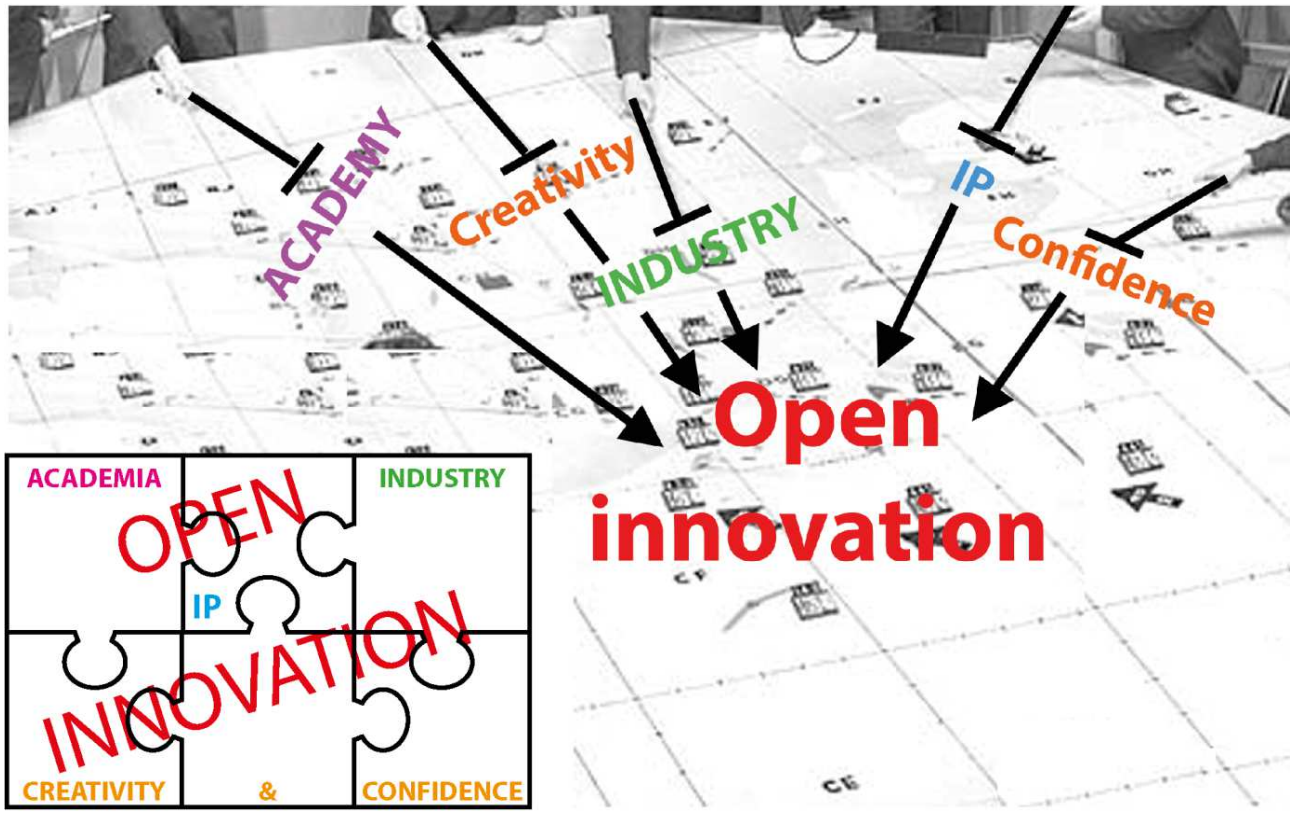

Fig. 5 Visual analytics support innovation through collaborative work
(top right part modified from www.subbrit.org.uk/rsg/sites/h/holmpton/index203.html).

3.2 WE MUST SEE FIRST

Visualization tools give a fast overview of a dataset. The website Gapminder (gapminder.org) is a perfect illustration of the visualization power over lists of data points. Simply download one of their excel files and compare it to the corresponding map. The first thing that is obvious from this exercise is how intuitive it is to use mapping tools. No effort is required to navigate through the data and find relevant information. On the contrary to the table that is far from being understandable and it would take an expert or some time to grasp the meaning of each data point. In the IP domain "patent properties makes patent analysis a field that fits a visual analytics approaches in general quite well, since the problem can neither be solved by human effort alone, nor fully automatically" [12] (Koch, 2012).

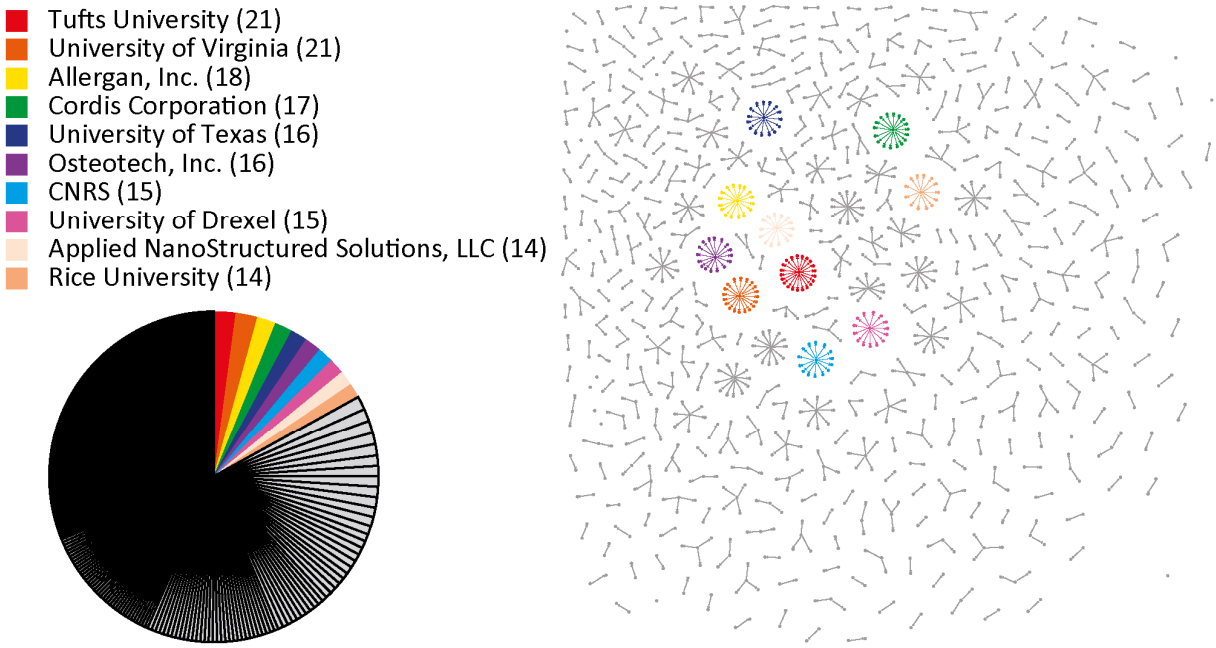

Fig. 6 Assignees/number of patents correlation in the field of natural polymers. Corresponding to 1000 patents in the field of natural polymers retrieved using a semantic search tool (query from [5] Rebouillat & M. Lapray, 2014). Pie-chart plotted in Excel and network plotted using software such as Sci² tool (the shape of each cluster does not carry any particular information and would be different every time the program would run on this data points). The density of each group is correlated to the relative assignee weight.

Figure 6 also illustrates the power of visualization using IP data. It is easy to see that even a simple pie chart, known for its numerous flaws [13] (Few, 2007), provides immediate better understanding of the patent assignees distribution than a list. Furthermore, it does not require any effort from the analyst to instantly see from the network map (right panel, Fig. 6) that there are few main players and the rest of the patents are evenly distributed between many institutions. If the demonstration is already so clear for such easy task, the reader can appreciate the potential of such tools for more complex problems (Fig. 7).

The rest of the review will assess how the IP domain is doing in terms of visualization tools and what the professional can use to ease the whole process.

	Patent titles
1	New fibre containing active elements, process for the preparation thereof and their uses, in particular in packaging
2	Fiber-assembled tissue constructs
3	Hybrid nanofibril matrices for use as tissue engineering devices
4	Immunoneutral silk-fiber-based medical devices
	…
999	PVA hydrogel
1000	Methods of bonding or modifying hydrogels using irradiation
1001	Methods of modifying hydrogels using irradiation

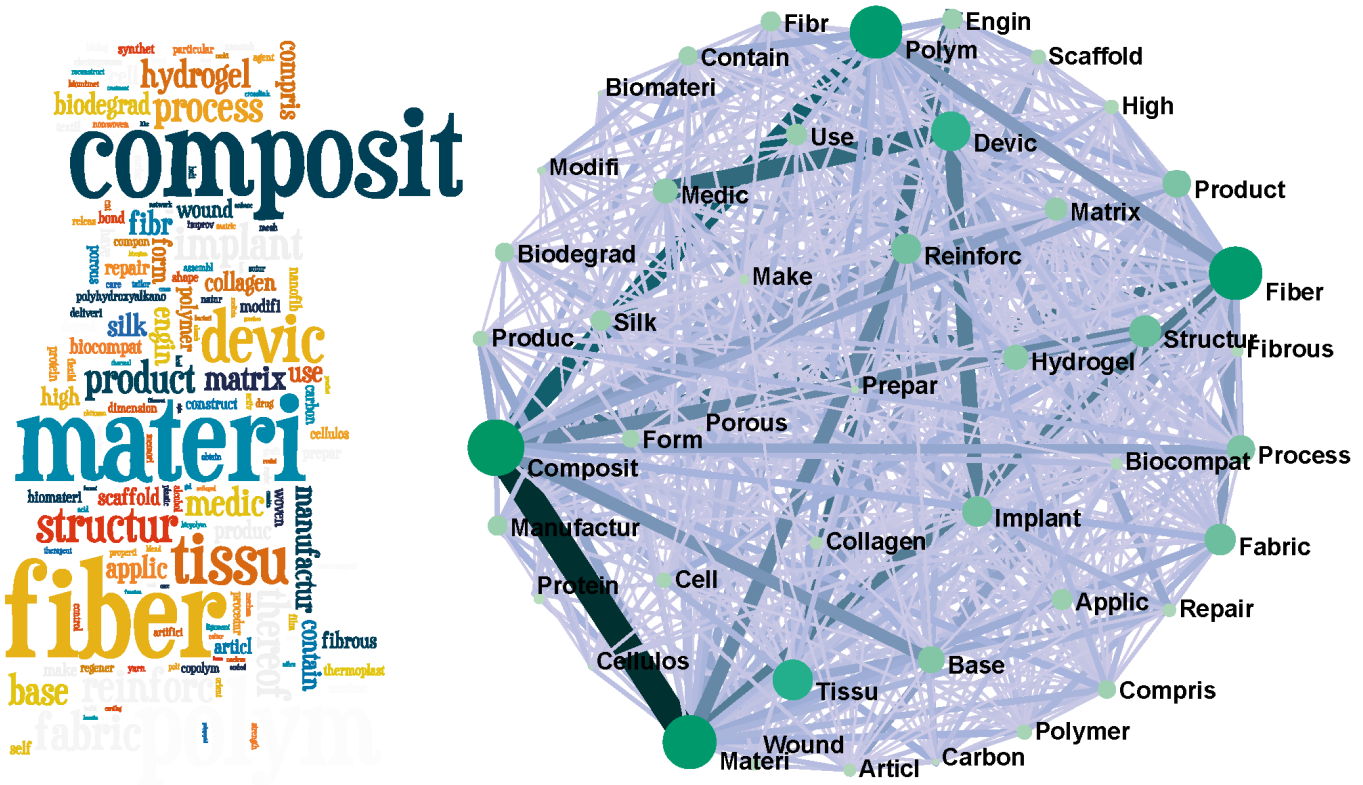

Fig. 7 Words co-occurrence network map. While a list of patent titles (from Fig. 6) does not carry any easily accessible information by itself, extracting and mapping words' weight as well as their relationships (right: directed network and Kamada-Kawai layout) can be of great interest. The shape, distance and placement of each node on the Kamada-Kawai graph layout do not carry any meaningful information (see Fig. 8). This layout has aesthetical benefits like all force-directed ones. It places nodes in a way that all edges are of more or less of equal length and minimizes the number of edge crossings. Here the weight of each node and edge is indicated by the size of each element as well as the darkness of the color. The titles were preprocessed [6] (Rebouillat & D. Lapray 2014) which explains their shorter appearance.

3.2.1 MATERIAL SELECTION, THE FUEL FOR CREATIVITY

Before any interpretation can start, there is an initial need to find the right information that will create the foundation of the whole process [6] (Rebouillat & D. Lapray 2014). The complexity of patent and non-patent literature prevents a straightforward retrieval of all relevant material and leads to tedious research processes. It is well known that in the case of patents, the language used differ between fields and the legal aspect of it adds to the apparent or real confusion pending on the agility and the experience of the operator. Therefore, analyzing patent data requires human expert judgment and cannot be completely automatized. The overwhelming large sets of results ask for new solutions to find relevant material [14] (Koch & Bosch, 2011b). Unfortunately, in the IP domain no application uses the visualization tools at any step of the process. The

retrieval has to be done in a more classic way using conventional Boolean queries or more sophisticated semantic search engines [6] (Rebouillat & D. Lapray, 2014).

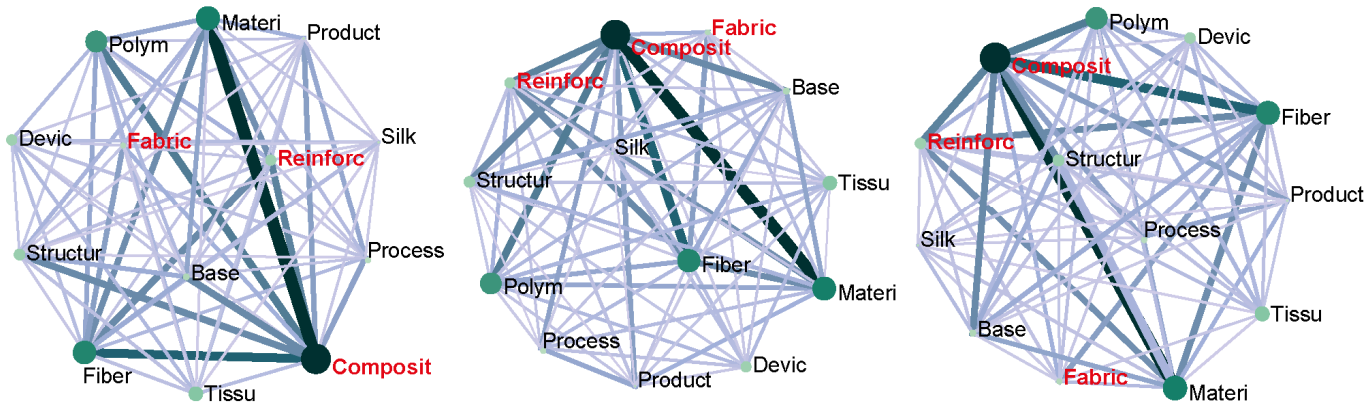

Fig. 8 Aesthetical tricks do not necessary carry relevant information. In this figure the same analysis was applied to the title of 1000 patents in the field of natural polymers. The 13 top nodes are displayed using the Kamada-Kawai layout. The three panels represent the exact same data on which the layout was reinitialized three consecutive times. Note that only the nodes, edges size and color are maintained between these three representations. These are the only relevant information, the rest being purely aesthetical.

3.2.1.1 VISUAL REPRESENTATION OF TEXTUAL QUERIES

The first significant move in the direction of visual query, and patent visualization in general, was initiated by the European project PatExpert (www.patexpert.org) and the development of PatViz [15] (Koch & Bosch, 2011a). This tool includes the possibility to link textual and visual query representation (Fig. 9). This potentially could be a powerful solution to the problem associated with the existing multiple Boolean syntaxes. The idea behind is relatively simple. The visual representation of a query is easier to write with less prior knowledge. Furthermore, the conversion to the different search engine syntaxes and languages would be done without intervention from the user [15] (Koch & Bosch, 2011a). This is unfortunately a pretty complex task and many difficulties stand before the developer.

Textual query:
(''biomimicry'' AND (''natural polymer'' OR energy OR pharma) AND NOT biotechnology)

Visual representation:

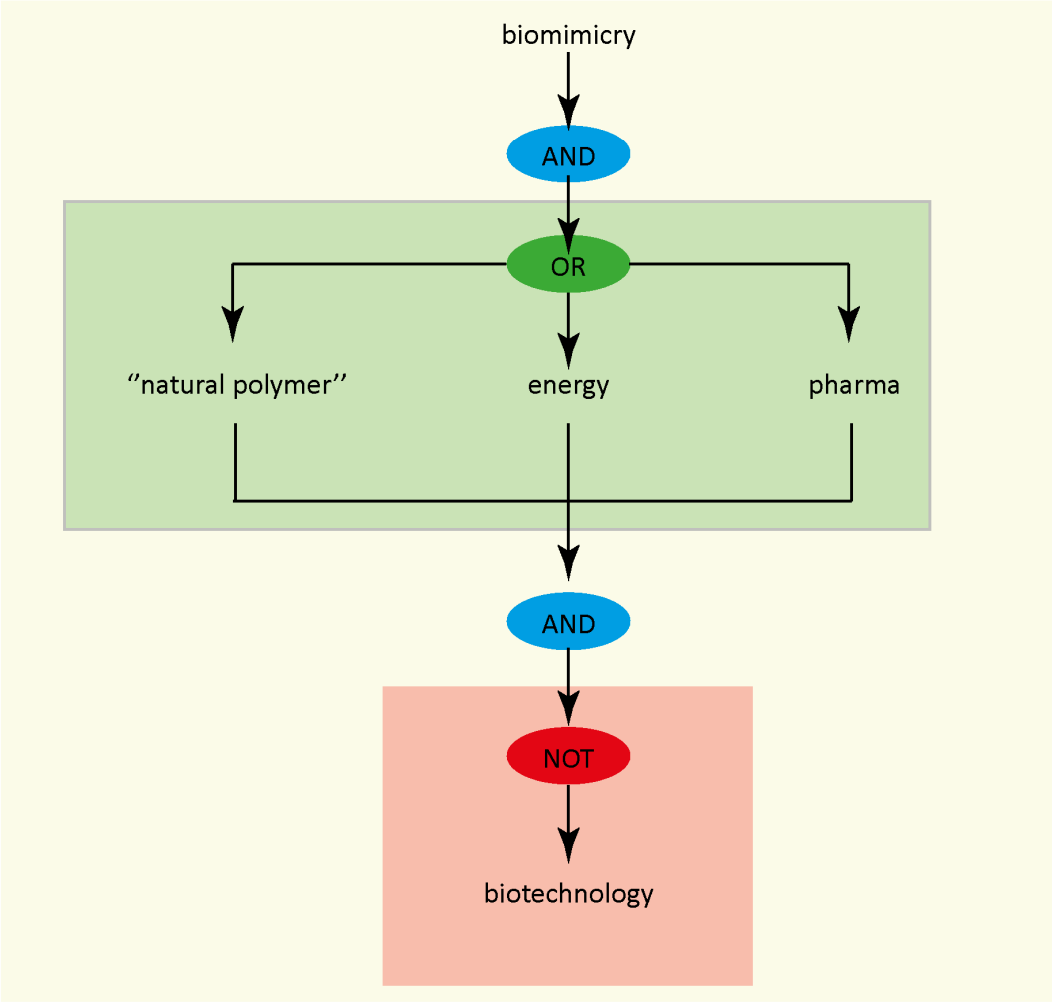

Fig. 9 Visual representation of a textual query. Query to retrieve the documents that contain the term "biomimicry" as well as any of the terms "natural polymer", "energy", or "pharma" but not the term "biotechnology".

Even if the initial query cannot be fully visualized, some software in the IP domain have made small but significant steps in the integration of visual tools in the initial selection of documents. InnovationQ analytics, for example, allows some sort of visualization to finely tune the field associated to an initial textual query. The main request in this case is a patent number or a simple term. The result is displayed on a map and the user can move around to visually modify the initial request. This is one of the strength of visualization, the possibility to cooperate with the system to reach a satisfactory level of confidence in the results. This kind of interactive way to work with the machine will be further explained in chapter 3.2.4.

Images are important part of patents and non-patent literature since they carry a lot of information [6] (Rebouillat & D. Lapray 2014). It is difficult to retrieve them and complex to use as initial. However, inspiration can come from other fields such as computer games. The Microsoft XBOX 360 Kinect sensor is a perfect example of image query and retrieval system. When a player makes a move to catch a ball the program does not analyze the actual gesture of the user. It compares the images taken during movement to a database, associates the query image to the according position and gives the adequate response. It might not be possible in a near future to use such technology in the IP domain due to the poor quality of patent images; however it would be of great interest for the whole field to change the way images are used.

3.2.1.2 FINDING A WAY THROUGH IMPERFECTION

It is not yet possible to use visualization tools at the query level as seen before, but mapping data points can be used afterwards in the process. The professional has two main options. The first one is to use "readymade tools" that propose to carry both the search and the visualization. It is the most convenient although not the most powerful and flexible solution. Without being exhaustive we can name few dominant programs such as Pantros IP (ip.com), Thomson Reuters IP solutions, IHS GoldFire solutions, Orbit IP (Questel-Orbit). These technologies offer interesting tools to retrieve data such as semantic or latent semantic analysis, avoiding some pitfalls associated with keywords retrieval systems [6] (Rebouillat & D. Lapray 2014). However, there visualization tools are limited to traditional charts and histograms. Orbit IP seems to propose more in terms of mapping. These tools are good to extract and save in readable format huge amount of patents that can be further analyzed using more specialized mapping tools. FreePatentOnline or Pubmed also propose solution to save search results in usable formats. Patent office search engines such as the United States Patent and Trademark Office (USPTO) or Esp@cenet from the European Patent Office (EPO) can also save research results for further analysis.

A variety of formats are available to aid data interchange such as CSV (comma-separated values), XML (Extensible Markup Language), Tab-delimited text files (UTF-8), XLS (Microsoft Excel file format) for the most commonly used. These formats allow the exchange between patent search solutions and visualization software such as the Science of Science (Sci^2) tool [16] (Sci2 Team, 2009), Pajek [17] (Batagelj & Mrvar, 2011), NAViGaTOR software (Network Analysis, Visualization and Graphing, Toronto, Canada), Cytoscape [18] (Shannon et al., 2003), Gephi (https://gephi.org/), etc. None of these open source software (for a more complete list of data mining and visualization software see [11] Keim et al., 2010) has been initially developed to work with patent data. One big advantage of visual analytics tools is their compatibility with most data format and with each other.

Before being able to visualize the results of document searches it is necessary to prepare the files for being used by different mapping software. Unfortunately, this cleaning phase can be tedious depending on the source of the retrieved documents. For most of them, it is necessary, for example, to correct assignee names that are very often misspelled. Visualization without this preliminary step would result in an under-representation of some clusters. When one wants to analyze word co-occurrence in titles, for example, passing the data through tokenization, stemming and stop-word removal will reduce the noise. When the data have been prepared, it is relatively easy to use different mapping solutions and facilitates the extraction of meaningful information.

3.2.2 GRAPH LAYOUT AND VISUALIZING ATTRIBUTES, CHOOSE WISELY!

"The adequacy of a visualization technique depends on the type of data to be shown as well as on the task to be accomplished with its help. Accordingly, not all graphical perspectives are well suited to display arbitrary types of information and different methods have to be employed for showing geospatial, temporal, hierarchical, network-based, etc. data." [15] (Koch & Bosch, 2011a). Furthermore, Diverse attributes and visualizations can offer different perspectives on the same data [19] (Pastrello et al., 2013). Many maps are in fact networks represented by a set of nodes, linked to each other with a set of edges [20] (Jurisica Lab, 2011). Depending on what type of network and/or layout is used, the same data points can look drastically different and therefore convey a lot or very little useful information (Fig. 10). It is therefore important, when mapping any kind of information to understand what each layout implies. Some arrangements are random and distances between data point do not carry any information by itself (Fig. 8). This is something to keep in mind when analysis such network and not be carried away on some illusionary interactions (see chapter 4 regarding visualization pitfalls).

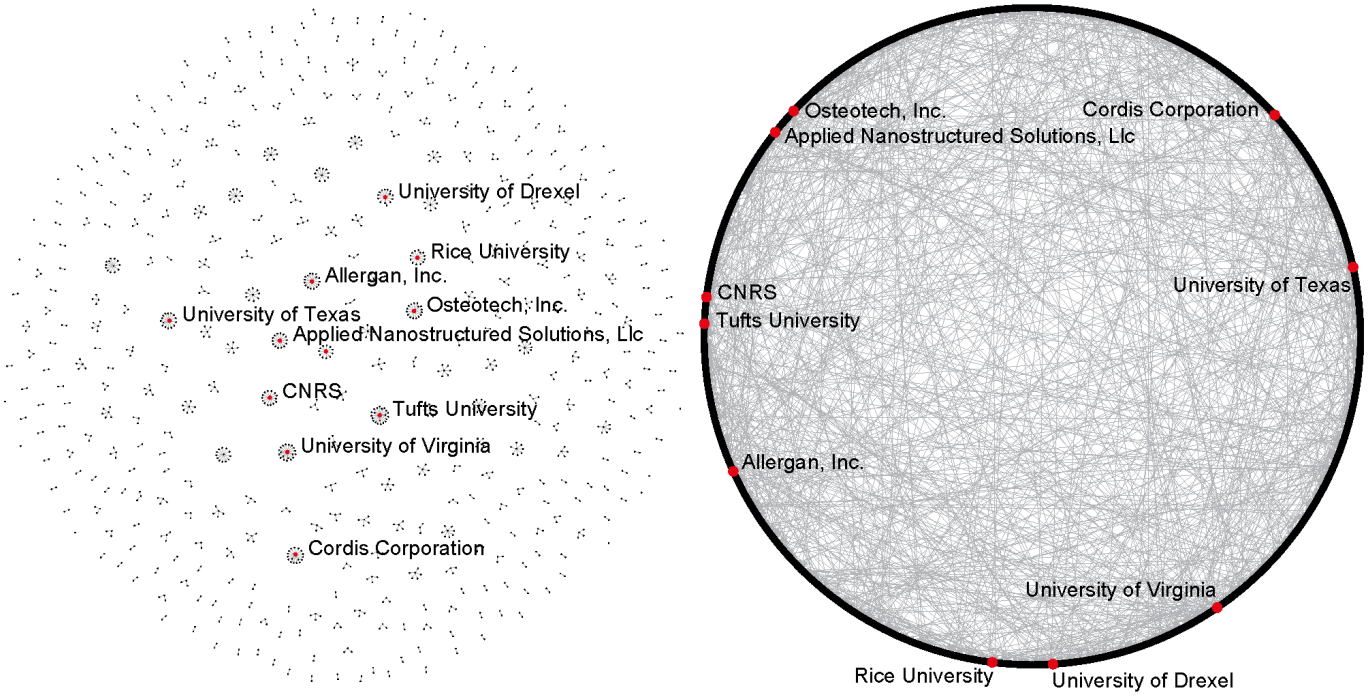

Fig. 10 Map layouts play a critical role. Directed network extraction using suitable visualization tool such as Sci^2. When mapping data (same as Fig. 3) it is important to choose a layout that conveys useful information.

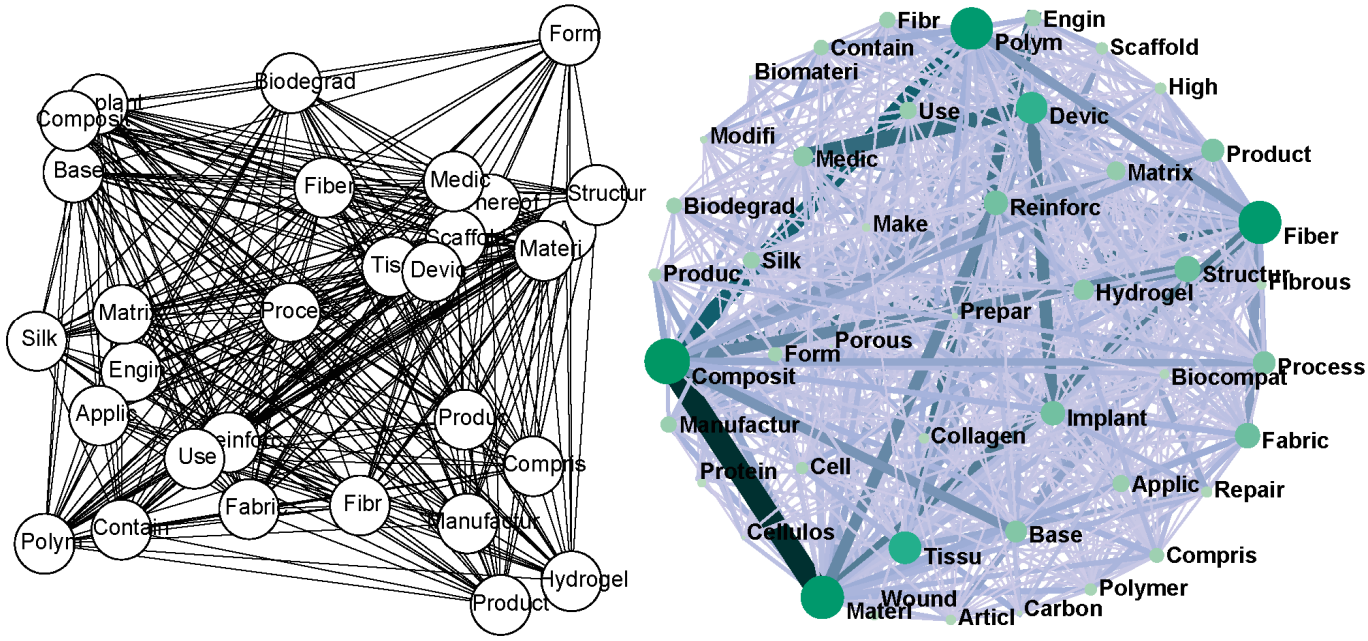

Fig. 11 Nodes and edges attributes are critical in the way a map is being perceived. These two maps represent the same information (from Fig. 7) however nodes and edges attributes are different. On the left panel all nodes and edges share the same attributes whereas on the right panel colour and size of nodes and edges are directly correlated to their respective weights.

The nodes and edges attributes are also of great importance. If we take Fig. 7 words co-occurrence network and we remove weight attributes from nodes and edges as well as color intensity/weight correlation the map loses complete readability (Fig. 11).

These options should therefore be considered seriously when using a mapping tool since the level of flexibility can improve the whole analytic process. It is also important to be able to navigate, select, focus, display, erase, interact with the visualization to explore the data [12] (Koch, 2012).

3.2.3 CLUSTERING VS CLASSIFICATION

When grouping data points together the word "clustering" appears very often. It is important to understand that any cluster of points does not necessary emerge from a "clustering method". There are two main ways to group data points: classification (or supervised method) and clustering (or unsupervised method). Many patent landscape analyses rely on the first method, classification. In this method the different classes are already known and the aim is to know how the data relate to those predefined categories. The International Patent Classification (IPC) co-classification map is a perfect classification illustration. The categories are known before any grouping and the aim is to attach every patent to its respective IPC group (Fig. 13, 14 and 19) [21] (Leydesdorff et al., 2012).

In the case of a clustering algorithm the data are grouped together using unknown initial parameters. Clustering methods try to group data points by finding whether there is some relationship between the objects without predetermined specification how records should be related together. Clustering is more used as an exploratory method [12], [22] (Koch, 2012; Yoon et al., 2013) to discover the unexpected [11] (Keim et al., 2010). For example the IP search tool Pantros IP gives a model of such method with its so called LSA clusters. The algorithm generates semantic clusters based on the documents at hand. These clusters will be different for every set of data presented to the algorithm. Classification and clustering methods have different objectives and use different ways to segment data.

3.2.4 HUMAN-MACHINE INTERACTION

We previously advocated for an increase of artificial intelligence in the IP search domain [6] (Rebouillat & D. Lapray, 2014). However, the experience of professionals acquired through years of data mining cannot be fully replaced by machines. This situation will not change in the near future but tools have to be introduced to help dealing with the quantity and diversity of the available information. The machine can exploit, calculate, clustered, etc., nevertheless so far only human interpretation can spot interesting patterns. A thought-provoking example is the observation made by scientists using Google™ Earth that cows like to align themselves with magnetic fields [23] (Begall et al., 2008). A correlation that machines alone could not have detected (see chapter 4 for the downside of this kind of studies).

Another point of difficulty is the confidence that should bind users and tools especially in the IP domain and it is "very difficult to build trust into mechanisms that cannot be fully controlled" [12] (Koch, 2012). This partly explains why Boolean retrieval systems are still favored over more sophisticated, therefore less transparent, methods. The professional faces a challenge because of the need to retrieve all relevant information. It is necessary to be sure of the retrieved material but it is not possible to control manually all the documents. The development of visualization technologies provides the perfect opportunity for a close human-machine interaction at every step of the process. Combining human and artificial intelligence is, to our opinion, a way to keep control over the course of actions necessary to find relevant data and to build trust. Having continuous visual feedbacks as proposed by PatViz [14] (Koch & Bosch, 2011b) and direct human-machine interaction to reach the optimal result is of great interest for innovation. A better control over the analytical process using continuous visual interactions is also a way to prevent the miss-reading of ready-made results. However, this necessitates great efforts of transparence from companies providing analytical services. The black-box concept cannot sustain in a domain where every step of the search has to be controlled because missing data is a "lost" discovery.

4 I SEE THINGS THAT DO NOT EXIST - THE VISUALIZATION PITFALLS

As mentioned previously, data observation using visual analytics offers endless possibilities to discover the undetected. However, the human brain is an expert in finding correlations between events or data points. Does that mean that what I see and interpret to be correlated really is? Coming back to the cows' alignment to magnetic fields mentioned in the previous chapter. After the first publication, other teams looked at this problem and could not replicate the data [24] (Cressey, 2011). It is not the purpose here to discuss the validity of the method or the quality of the science behind this study; rather to highlight the fact that observation, and in this case visualization, is a tool to handle with great caution.

The danger is to see apparent relationships or proximities that do not actually exist [10] (Bollier, 2010). A good example of such visual trick comes from the London Underground map [25] (Sharif, 2013) that has been shown to distort real distance between points. On this map the Lancaster Gate station and Paddington Station look distant enough to take the metro. In reality these two stations are at walking distance from each other's. When using visual data analysis it is crucial to be sure that what we see does exist and is not an optical illusion. One way to avoid this "dangerous" pitfall is to use several views of the same data. A real relationship should be apparent not only on one map. The integration in PatViz of multiple views in one

panel representing the data at hands makes a lot of sense to not be “tricked” by non-existent data proximities or distances. Backing up visual relationship with actual numbers is also a way to avoid mistakes of interpretation.

5 CASE STUDIES

In order to illustrate the use of visualization tools in the IP domain, patents from three different applications of the biomimicry domain were retrieved. The idea of this paper being to continue the work previously initiated [3], [5], [6] (Rebouillat & M. Lapray, 2014; Rebouillat & D. Lapray, 2014; Rebouillat, 2013). Two main software using semantic technology were used, system P and X (in red and green in the following figures respectively). Three sections of Rebouillat and Lapray review on biomimicking [5] (Rebouillat & M. Lapray, 2014) were used independently to generate the queries. We retrieved around 1000 patents for each selected domain and kept for further analysis (around 6000 documents). Furthermore, patents’ information (e.g. document number, IPC, priority date and assignees) were retrieved using only system X for consistency. Documents were saved as csv files and cleaned from mistakes such as assignees misspelling, missing elements or wrongly organized cells. The following maps were made using open source software such as Sci^2 tool, Pajek, Cytoscape, Gephi and VOSViewer.

5.1 NATURAL POLYMERS

5.1.1 PATENTS’ TEMPORAL DISTRIBUTION

Files containing the patents, generated from the “natural polymer” section of Rebouillat & M. Lapray, 2014 [5], were imported in the ad hoc tool and temporal bar graph visualizations using priority dates were generated (Fig. 12). The graph shows a divergence in the temporal distribution of the retrieved patents for the period before 2004. In this particular case we choose to assign a similar weight to each patent.

Such information, applied to a larger number of patents, such as 40000, is very important to understand the advent of technology and its disruptive nature. Out of curiosity one has noticed that the patent lists may differ considerably from system X vs. P and still the invention pace and growth rate remains concomitant.

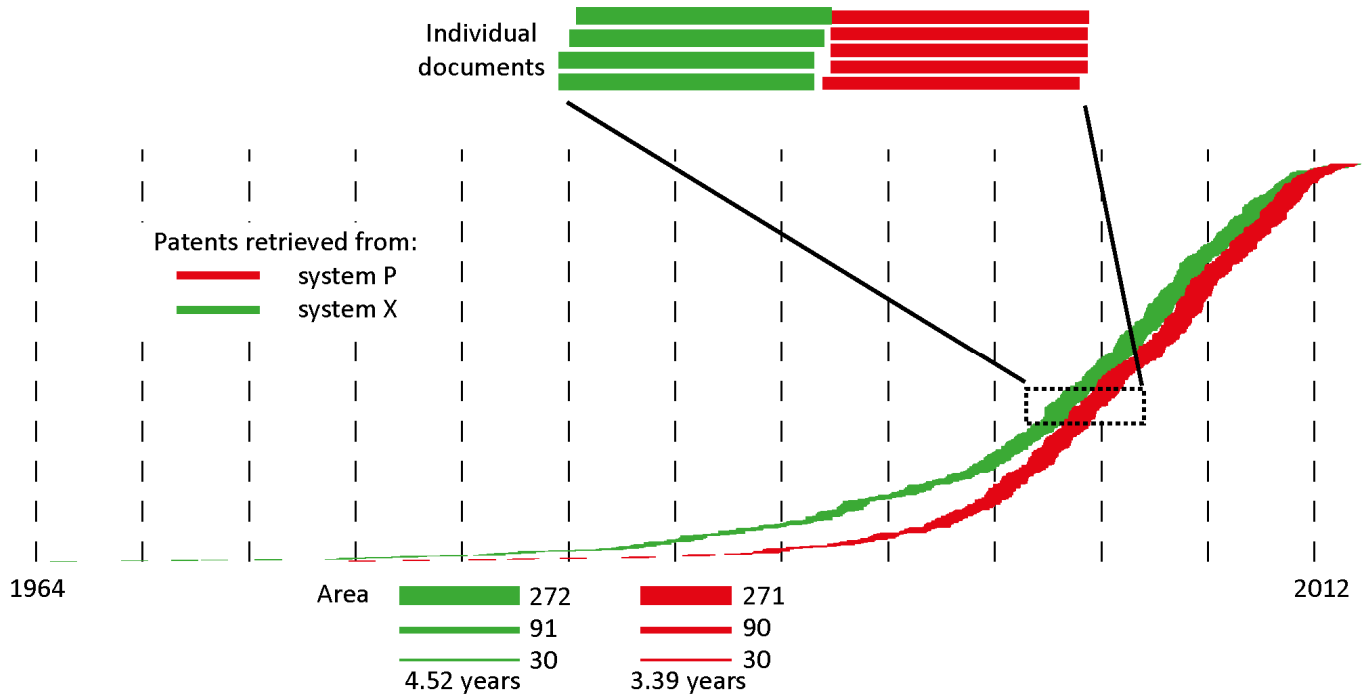

Fig. 12 Patents’ temporal distribution retrieved by the two semantic systems. In this graph the size of each document’s representation is similar and the number of documents is represented in terms of area. Figure produced with tool capable of cumulative graphing.

5.1.2 IPC CO-CLASSIFICATION ANALYSIS

The csv files previously imported were re-used and directed networks were extracted (IPC and publication number as source and target respectively). The networks were then saved as net files and imported in an ad hoc network analysis and visualization system such as Pajek. Fruchterman-Reingold maps were produced and exported into VOSViewer to create density views (Fig. 13) [21] (Leydesdorff et al., 2012). The Fruchterman-Reingold algorithm belongs to the force-directed ones and the placement of each node is stable when the system reaches equilibrium at minimum energy. As was seen in the temporal graph, the retrieved documents cover different areas highlighted, for example, by the over-representation of the "organic macromolecular…" class of system P compare to system X.

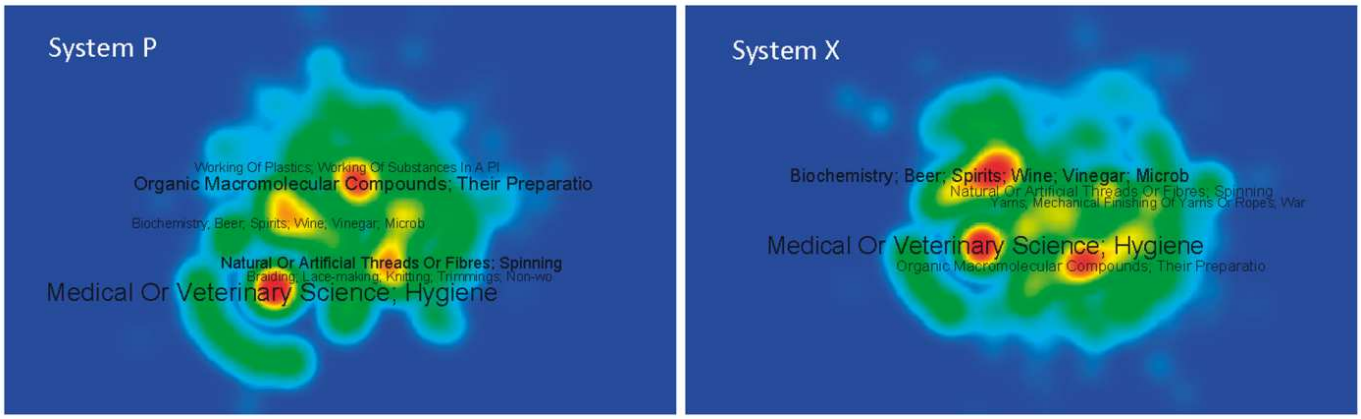

Fig. 13 IPC (3rd digit level) co-classification density views. These maps were generated using VOSViewer or alike.

5.1.3 ASSIGNEES DISTRIBUTION

Directed networks were extracted from the files (assignee and publication number as source and target respectively). Visualizations were then generated with display tools such as the Sci^2 tool or alike (Fig. 14). This layout minimizes edge intersections and randomly assigned edges size, e.g. in these graphs only the density of point carries information. From the 12 main players retrieved by the two systems only 3 were common to both sets but with different weights (highlighted in blue).

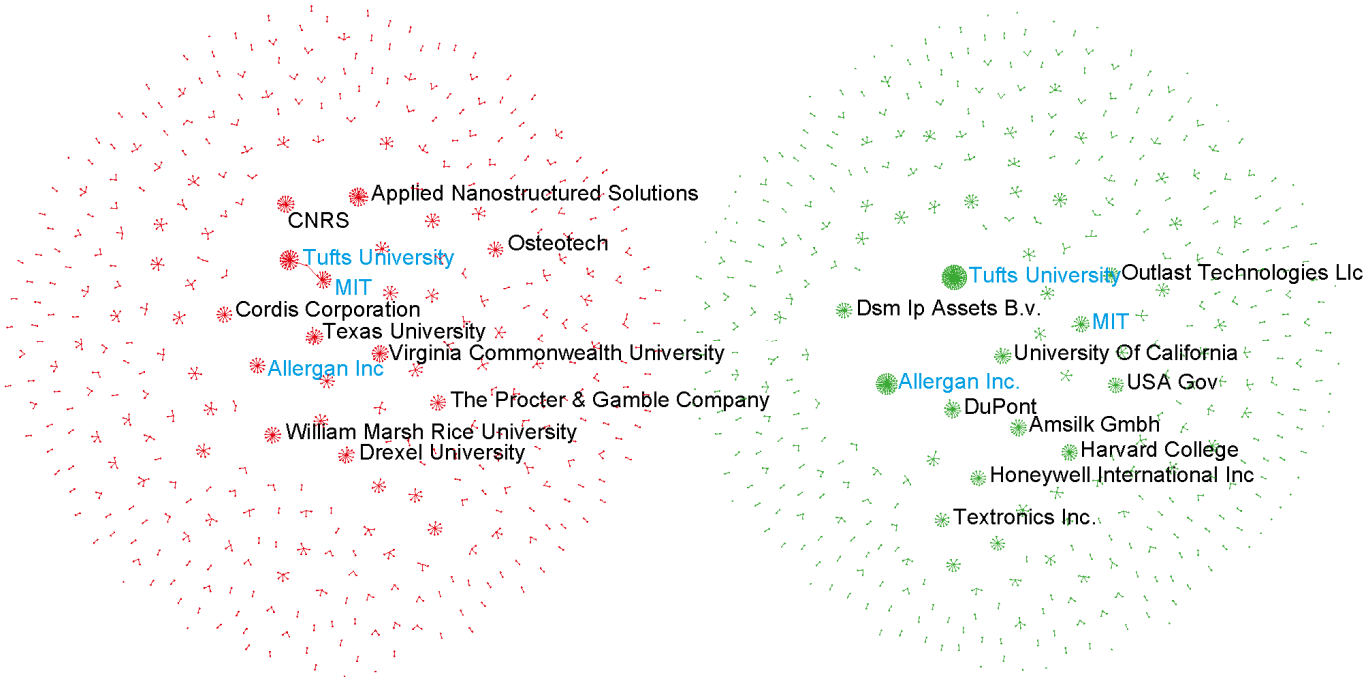

Fig. 14 Assignees distribution in the retrieved patents. The 12 main players are indicated (blue: common to the two systems).

5.2 ENERGY

5.2.1 PATENTS' TEMPORAL DISTRIBUTION

Files containing the patents, generated from the "energy" section of Rebouillat & M. Lapray, 2014 [5], were imported in the ad hoc tool and temporal bar graph visualizations using priority dates were generated (Fig. 15). The graph shows high temporal distribution similarity between the retrieved patents.

Such information, applied to a larger number of patents, such as 40000, is very important to understand the advent of technology and its disruptive nature. Out of curiosity one has noticed that the patent lists may differ considerably from system X (green) vs. P (red) and still the invention pace and growth rate remains concomitant.

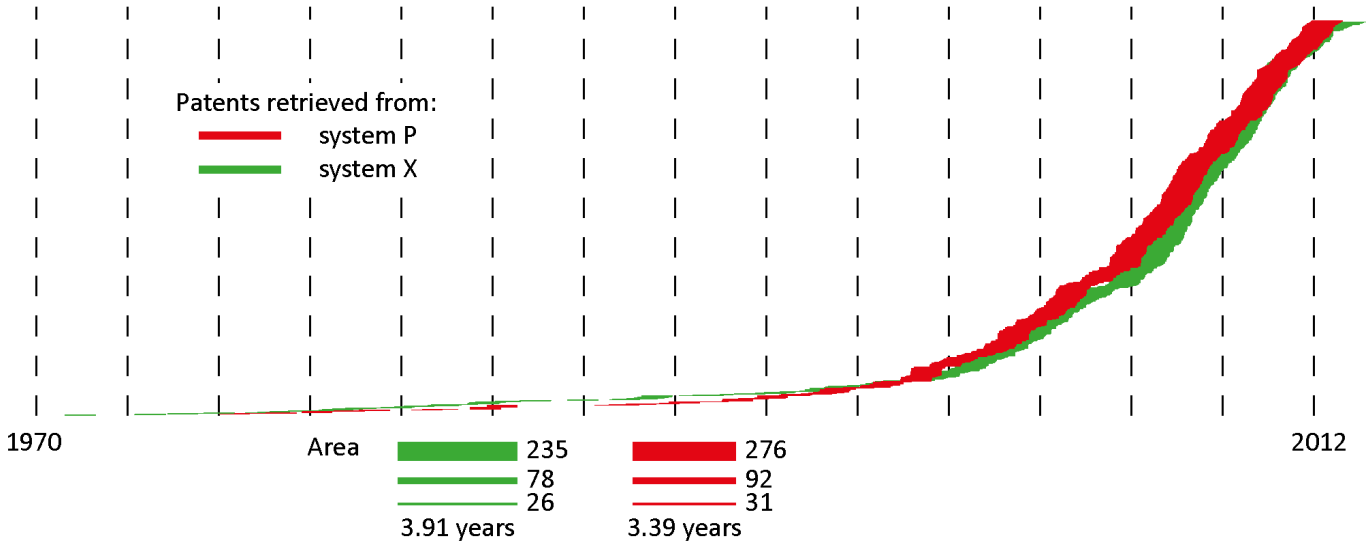

Fig. 15 Patents' temporal distribution for the two semantic systems. Graph produced with tool capable of cumulitative graphing.

5.2.2 IPC CO-CLASSIFICATION ANALYSIS

Directed networks were extracted from the files (IPC and publication number as source and target respectively). Force directed network visualizations were used to map the IPC co-classification (Fig. 16). Nodes were force positioned in order to get a clear picture of the main components of these networks. As any force-directed algorithm the aim is to point in a way that all edges are of more or less of equal length and to minimize the number of edge crossings. As seen with the temporal graph, the retrieved documents cover similar IPC areas.

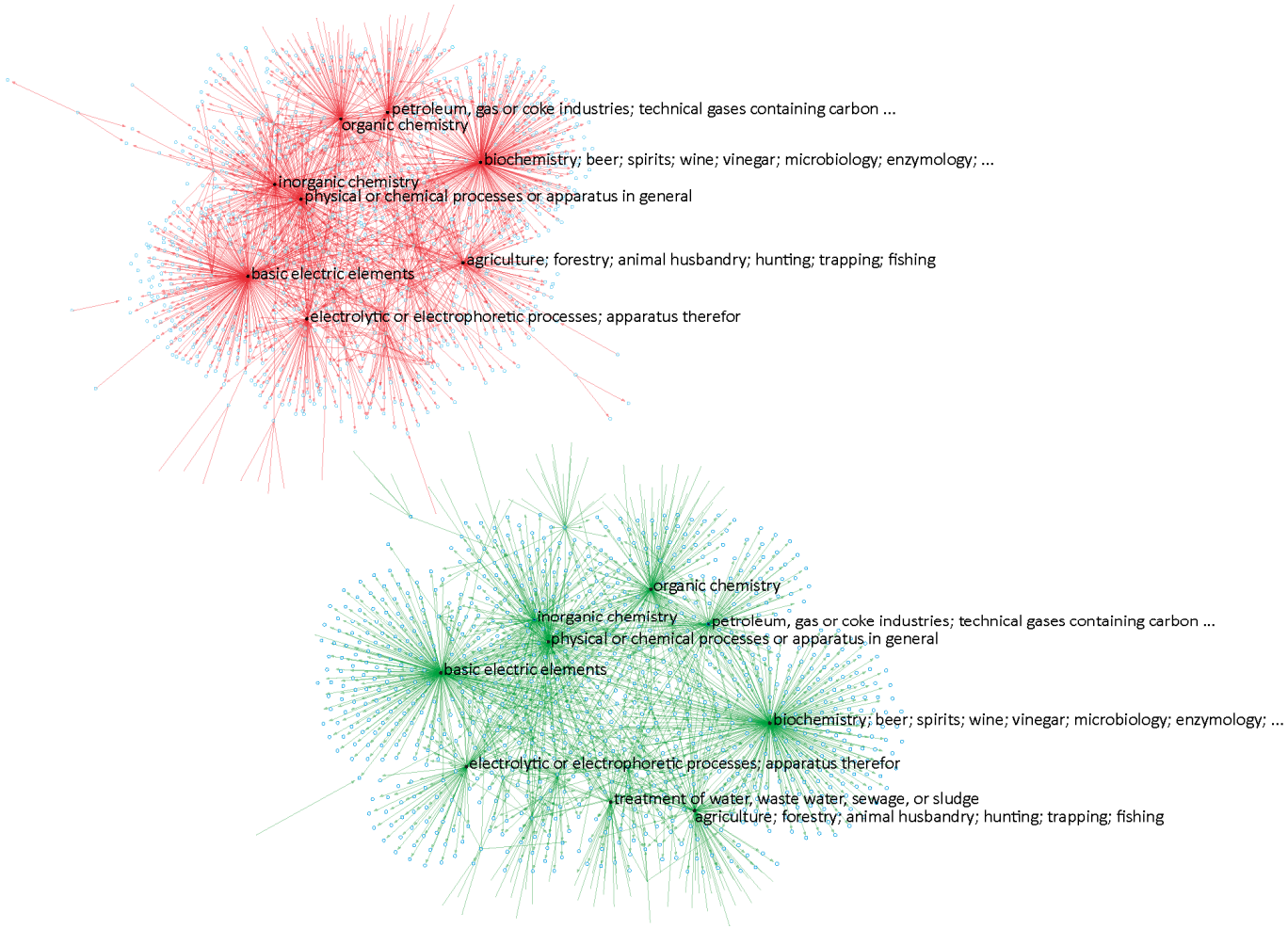

Fig. 16 IPC (3rd digit level) co-classification force directed visualization. Networks generated by ad hoc software such as Sci² tool.

5.2.3 ASSIGNEES DISTRIBUTION

Directed networks were extracted from the files (assignee and publication number as source and target respectively). Visualizations were then generated with display tools such as Cytoscape or alike, using the degree sorted cycle to layout the graph (Fig. 17). Despite the similarity in the covered IPC areas and temporal distribution, there are significant differences regarding the assignees weights.

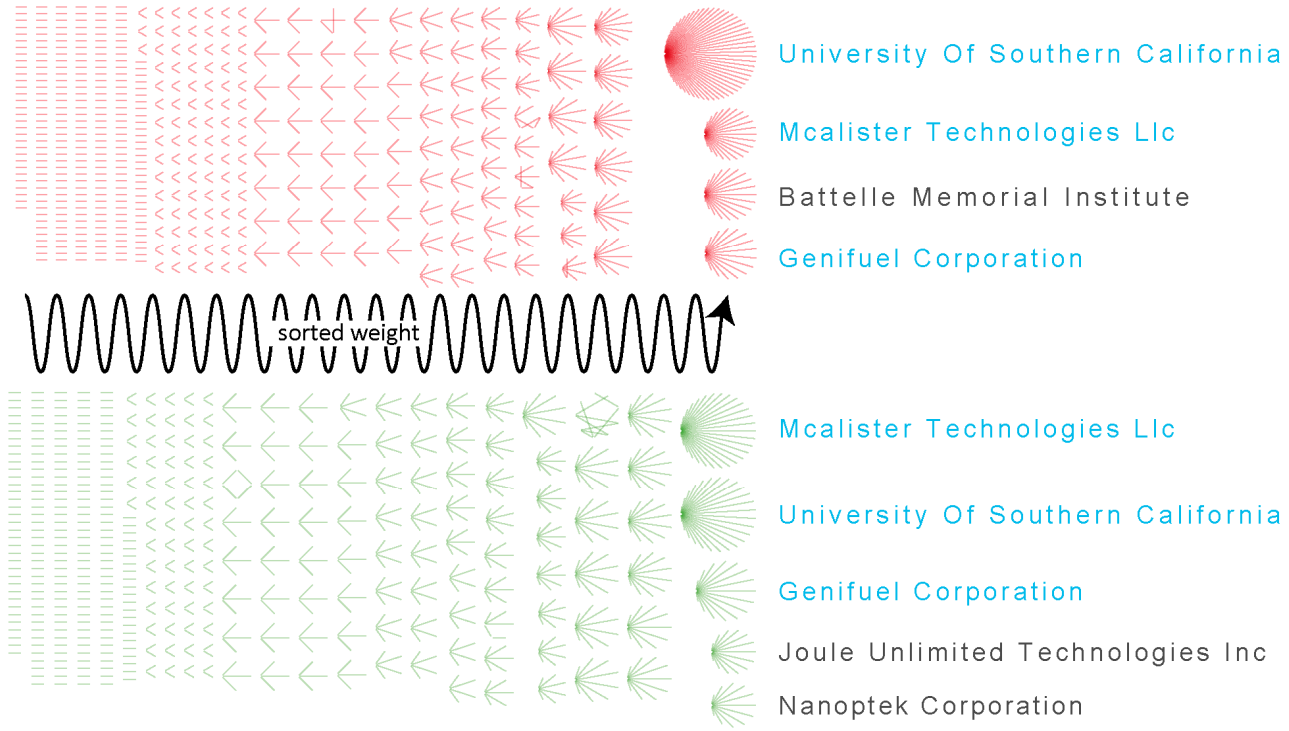

Fig. 17 Assignees distribution in the retrieved patents. The main players are indicated (blue: common to the two systems). Drawing made with graphing software such as Cytoscape or alike.

5.3 PHARMA

5.3.1 PATENTS' TEMPORAL DISTRIBUTION

Files containing the patents, generated from the "pharma" section of Rebouillat & M. Lapray, 2014 [5], were imported in the ad hoc tool and temporal bar graph visualizations using priority dates were created (Fig. 18). The graph shows a high temporal distribution similarity between the retrieved patents.

Such information, applied to a larger number of patents, such as 40000, is very important to understand the advent of technology and its disruptive nature. Out of curiosity one has noticed that the patent lists may differ considerably from system X vs. P and still the invention pace and growth rate remains concomitant.

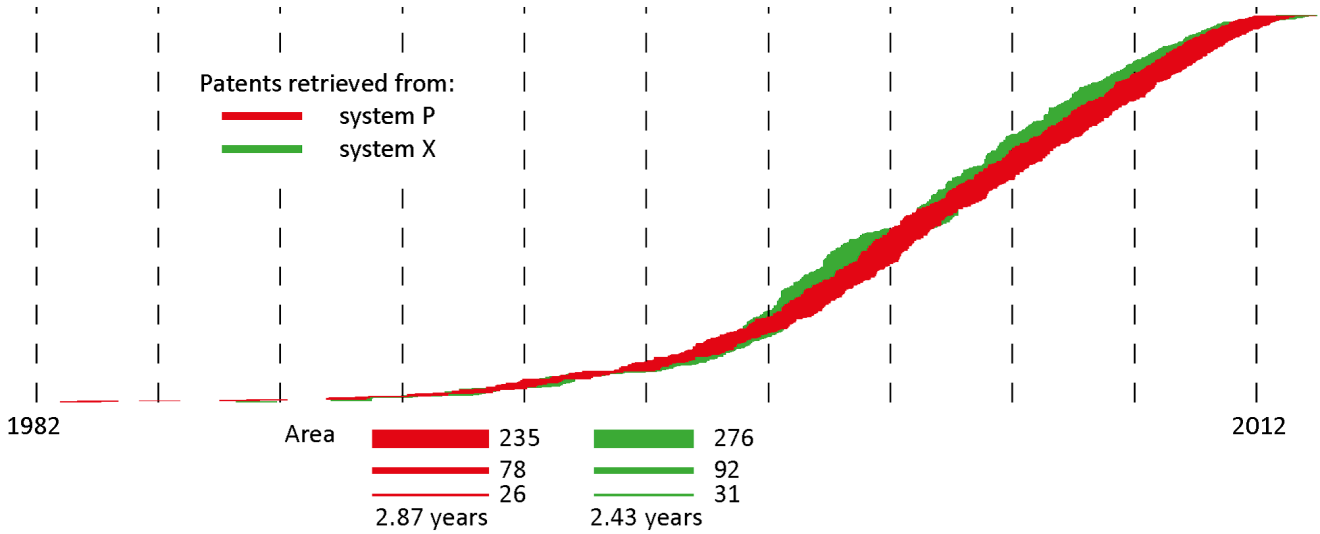

Fig. 18 Temporal distribution of the patents on pharma retrieved from the two semantic systems. Graph produced with tool capable of cumulative graphing.

5.3.2 IPC CO-CLASSIFICATION ANALYSIS

Directed networks were extracted from the files (IPC and publication number as source and target respectively). Suitable layout was used to visualize the IPC co-classification (Fig. 19). The retrieved documents show important differences in the IPC area distribution. For example, the percentage of patents attached to the classification code A61 (medical or veterinary…) represent 55.8 and 68.3 % of system P (red) and X (green) retrieved patents respectively. For the class C12 (Biochemistry, beer…), the difference between the two sets of patents is more important with 30.4 and 63.8 % documents classified in this category for system P and X respectively (percentages retrieved using suitable retrieval tool such as Cytoscape or alike).

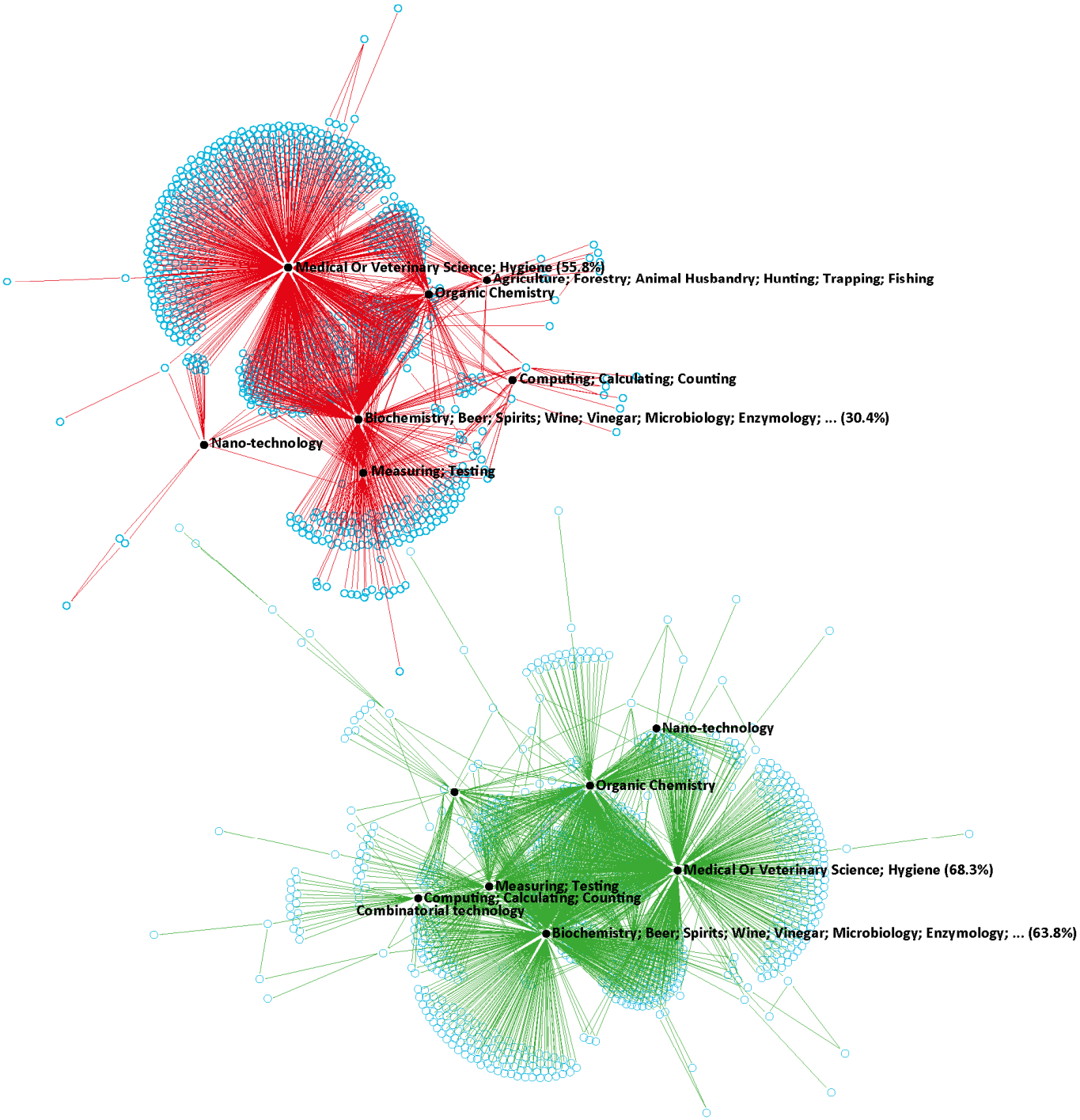

Fig. 19 IPC (3rd digit level) co-classification force directed visualization. IPC A60 and C12 are the most represented areas in the retrieved documents. In these graphs edges' length were generated by the GEM ("Generalized Expectation Maximization") algorithm to minimize intersections for aesthetical purposes. Maps generated with ad hoc software such as Sci² tool.

5.3.3 ASSIGNEES DISTRIBUTION

Directed networks were extracted from files (assignee and publication number as source and target respectively). Visualizations were then generated with display tool such as Gephi or alike, using a dual circle to layout the graph. Nodes were ordered by out-degree and 10 upper order nodes were placed out in a counter-clockwise manner for clarity (Fig. 20). The distribution of assignees illustrates very well the dissimilarity of both retrieved set of patents.

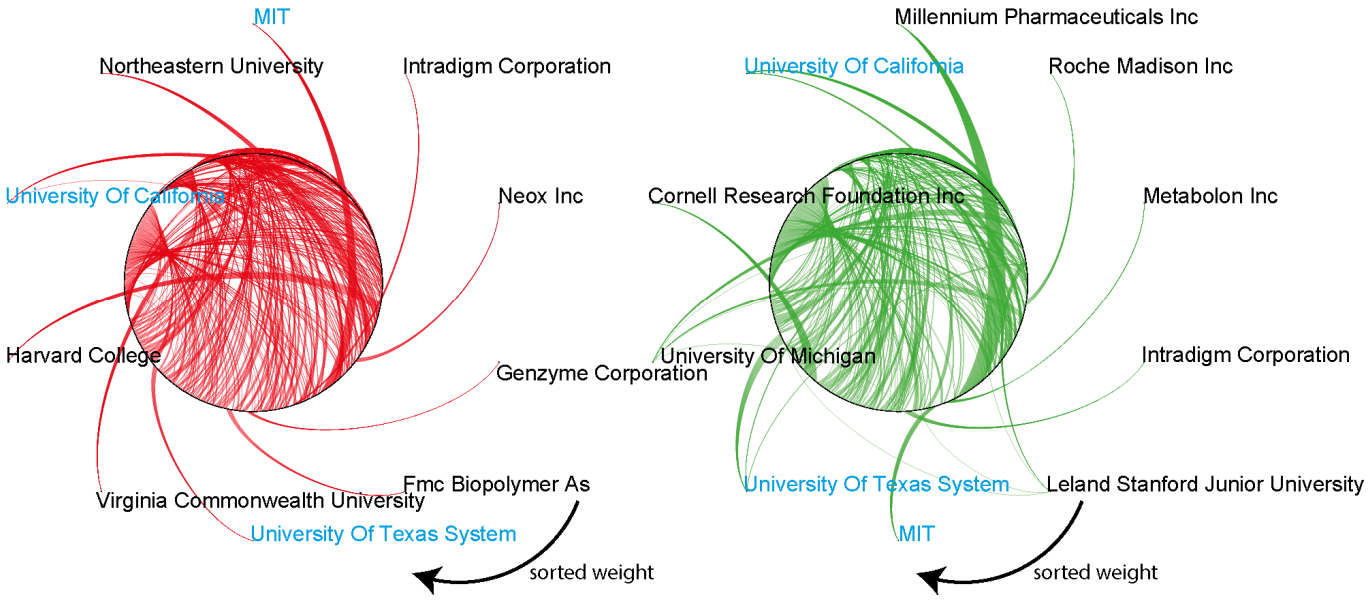

Fig. 20 Assignees distribution in the retrieved patents. The 10 main players are indicated (blue: common to the two systems). Layouts were generated using suitable software such as Gephi.

5.4 WHAT DID WE LEARN FROM THOSE MAPS?

All the information previously displayed would have necessitated expert knowledge to extract them from bar charts or tables. In the situation of more complex searches it would not be possible to go without a first visualization of the data. Several information conclusions can be drawn from these relatively easy case studies. First of all, visual mapping of information is a powerful tool that can ease the task of professionals to quickly detect relevant patterns. Mapping tools exist and can be used with IP material. Not initially developed to use such data, all these software can load and display multiple networks with different degrees of flexibility and interaction. Finally, semantic technologies offer great potentials but need to be used with a little bit of caution. The retrieved sets of documents might differ from one tool to another and it is therefore advisable to not restrict the search to only one. The use of visualization tools allowed a quick review of the available data; it is equally crucial to know what stands behind the different layout algorithm to be able to assess what is relevant and what is purely aesthetical.

Up to 20 to 40 000 patents were analyzed using these tools in a matter of a day, this is basically impossible to perform from a human capability standpoint. Nonetheless, the old adage “garbage in, garbage out” has to be taken seriously even before considering the retrieval of patent of interest. Regardless of the fine tuning quality, there is barely any integrated patent search engine today that can provide such diverse outputs in terms of trends and analysis. The expert can certainly gain time and move faster towards some directions, the neophytes might and will struggle much more. Therefore, the need to improve the IP visual analytics is urgently stressed.

6 CONCLUSION

Visualization tools to analyze IP material without having to acquire new and complex knowledge is at reach. Despite the lack of proper technologies fully focused on the visual analysis of IP data, professionals have plenty of solutions at hand. Inspiration can come from any field of research that uses mapping technologies and tools directly apply onto IP documents. Most of these technologies are based on the Cyberinfrastructure Shell (CIShell) [26] (Herr et al., 2006) and Open Services Gateway Initiative Framework (OSGi) industry standard that support integration of new and existing algorithms [27] (Börner,

2011). Plugins can easily be implemented to create and share among the community new ideas and solutions to display and extract information. The professionals wishing to take a chance on this path are not necessary bound to commercial solutions and can adapt their tools to special needs. Innovation can benefit from a community effort to tackle the Big Data challenge that no industry can ignore.

Innovation and inventions can largely be transformed by visualization, images and motions in general; visual analytics being the next step. The patent language barriers can likely be reduced from new semantic analysis converting thousands of words in new pictures and thousands of pictures in new concepts; 3D animations and virtual comparisons with real situations, including patent images and schematics, hold great potential for fast and disruptive open innovation.

All views are worth the climb in this domain, no one can afford ignoring the advent of "Big Data" as a source of inspiration beyond human conception and pace. Mastering the inputs to optimize the outputs quality is even more important given the unlimited numbers of direct or indirect references that can be used.

The multidisciplinary dimensions of a number of projects can be simplified given the opportunity to relate visual to textual and vice versa.

The next chapter of this series is focusing on the open and disruptive innovation metrics.

ACKNOWLEDGEMENTS

The authors thank:

Mirosława Lapray, for private conversations yielding useful improvements.

More about the authors:

Serge Rebouillat [Rebooya], Dr. Ing., Dr-ès-Sciences, Certified Prof., Ind. Energetic, Chem/Bio Eng., Rheology, IP/Mediation/Innovation & Strategy.

Damien Lapray, Ph.D., Neural Network Analysis.

REFERENCES

[1] J. Dyer, H. Gregersen, and C. Christensen, "The innovator's DNA," *Harv. Bus. Rev.*, 2009.
[2] J. West and M. Bogers, "Profiting from external innovation: a review of research on open innovation," *Soc. Sci. Res. Netw.*, 2011.
[3] S. Rebouillat, "A Science & Business Equation for Collaborative Corporate Innovation . Business Strategy, IP Strategy, R&D Strategy: an all-in-one Business Model. A review with a Bio-Technology & Green Chemistry Focus," *Int. J. Innov. Appl. Stud.*, vol. 4, no. 1, pp. 1–19, 2013.
[4] V. Marx, "The Big Challenges of Big Data," *Nature*, vol. 498, 2013.
[5] S. Rebouillat and M. Lapray, "Bio-inspired and Bio-inspiration : a Disruptive Innovation Opportunity or a Matter of " Semantic"? A Review of a "stronger than logic" Creative Path based on Curiosity and Confidence (4C22C©)," *Int. J. Innov. Appl. Stud.*, vol. 6, no. 3, pp. 299–325, 2014.
[6] S. Rebouillat and D. Lapray, "A Review assessing the "used in the art" Intellectual Property Search Methods and the Innovation Impact therewith," *Int. J. Innov. Appl. Stud.*, vol. 5, no. 3, pp. 160–191, 2014.
[7] H. Chesbrough, "Open innovation: a new paradigm for understanding industrial innovation," *Open Innov. Res. a new Paradig.*, 2006.
[8] Intel, "Big Data Visualization : Turning Big Data Into Big Insights," March, 2013.
[9] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," 2008. [Online]. Available: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.
[10] D. Bollier, "The promise and peril of big data," ISBN: 0-89843-516-1. 2010.
[11] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, "Mastering The Information Age-Solving Problems with Visual Analytics," ISBN 978-3-905673-77-7, 2010.
[12] S. Koch, "Visual search and analysis of documents in the intellectual property domain," Universität Stuttgart, 2012.
[13] S. Few, "Save the pies for dessert," *Vis. Bus. Intell. Newsl.*, pp. 1–14, 2007.
[14] S. Koch and H. Bosch, "Iterative integration of visual insights during scalable patent search and analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 5, pp. 557–569, Jun. 2011.

[15] S. Koch and H. Bosch, “From Static Textual Display of Patents to Graphical Interactions,” in *Current Challenges in Patent Information Retrieval*, vol. 29, M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 217–235, 2011.
[16] Sci2 Team, “Science of Science (Sci2) Tool,” 2009. [Online]. Available: https://sci2.cns.iu.edu.
[17] V. Batagelj and A. Mrvar, “Pajek: Program for Analysis and Visualization of Large Networks,” *Version 0.71*, 2011. [Online]. Available: http://vlado.fmf.uni-lj.si/pub/networks/Pajek/doc/pajekman.pdf.
[18] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape: a software environment for integrated models of biomolecular interaction networks.,” *Genome Res.*, vol. 13, no. 11, pp. 2498–504, Nov. 2003.
[19] C. Pastrello, D. Otasek, K. Fortney, G. Agapito, M. Cannataro, E. Shirdel, and I. Jurisica, “Visual data mining of biological networks: one size does not fit all.,” *PLoS Comput. Biol.*, vol. 9, no. 1, p. e1002833, Jan. 2013.
[20] Jurisica Lab, “NAVIGaTOR,” 2011. [Online]. Available: http://ophid.utoronto.ca/navigator/documentation/v2.1.12/.
[21] L. Leydesdorff, D. Kushnir, and I. Rafols, “Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC),” *Scientometrics*, pp. 1–23, 2012.
[22] J. Yoon, H. Park, and K. Kim, “Identifying technological competition trends for R&D planning using dynamic patent maps: SAO-based content analysis,” *Scientometrics*, vol. 94, no. 1, pp. 313–331, Sep. 2013.
[23] S. Begall, J. Červený, J. Neef, O. Vojtech, and H. Burda, “Magnetic alignment in grazing and resting cattle and deer,” *PNAS*, 2008.
[24] D. Cressey, “The mystery of the magnetic cows,” 2011. [Online]. Available: http://www.nature.com/news/the-mystery-of-the-magnetic-cows-1.9350.
[25] S. Sharif, “The Power and Danger of Data Visualization,” *LunaMetrics*, 2013. [Online]. Available: http://www.lunametrics.com/blog/2013/02/04/power-danger-data-visualization/#sr=mjgfboedpef.uvncms.dpn&m=r&cp=(sfgfssbm)&ct=/qptu/43853129982/uif-qpxfs-boe-ebohfs-pg-ebub-wjtvbmj.
[26] B. W. Herr, W. Huang, S. Penumarthy, and K. Börner, “Designing highly flexible and usable cyberinfrastructures for convergence.,” *Ann. N. Y. Acad. Sci.*, vol. 1093, pp. 161–79, Dec. 2006.
[27] K. Börner, “Plug-and-play macroscopes,” *Commun. ACM*, vol. 54, no. 3, p. 60, Mar. 2011.