Automated Patent Mapping: a source of IP-based intelligence?

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE
in Management of Technology
Faculty of Technology, Policy and Management
by
Jaap van de Vorst
Student number: 4453360

To be defended in public on October 12th, 2018

Graduation committee
Chairperson : Dr. ir. S. W. Cunningham, Multi-Actor Systems
First Supervisor : Dr. L. Hartmann, Delft Centre for Entrepreneurship
Second Supervisor : Dr. ir. S. W. Cunningham, Multi-Actor Systems
External Supervisor : Dr. ir. E. Balidemaj, Philips Intellectual Property & Standards
Acknowledgements

I would like to start this report with a word of thanks to everyone that has been a support for me during the execution of this research and the writing of this accompanying report. Without the supportive and intelligent people I was surrounded with during the execution of my research, it would have never reached the quality it has today. In the first place my thank goes to my supervisors from Delft Technical University. I would like to thank dr. Dap Hartmann for his guidance through the entire process, from setting up a thesis research until the final dot in this report. I could count on Dap for clear advice whenever I got stuck. This very same holds for dr. ir. Scott Cunningham, who was of great contribution by the guidance he provided during the regular meetings we had. Thanks to Scott I also experienced a highly steep learning curve in the programming language he exposed me to; Python. I am very thankful that I have been able to learn the basics of this language and to apply it for my research. Thanks also goes to Gerard van der Ligt, who offered me the possibility to execute my research at Philips IP&S department. On behalf of Philips, Gerard provided me with everything I needed in the execution of my research, including very nice colleagues. Of course special thanks goes to my company supervisor: dr. ir. Edmond Balidemaj, who helped me through all the facets of my research, especially with the extensive expertise he beholds in the field of patent analytics. Further, I would also like to thank Frans Obers and Thomas Tsatsarelis, both IP-analysts at Philips, for their participation in the experiment I exposed them to. Also big thanks goes also to friends and family, in particular my parents for their unconditional belief in me and two of my close friends, Bob Driessen and Rogier Versteeg who were of great help when I was dealing with the, respectively, complicated mathematics and programming related to text data visualisation. Last but not least special thanks goes to my girlfriend, who has been there for me through the entire process despite my long days of work and my sometimes obsessive behaviour in regards to the project. Thank you.

Nijmegen, September 2018

Jaap van de Vorst
Summary

Companies, now more than ever, recognize the value of patent analysis in contribution to their competitive and business intelligence. In highly competitive technology oriented industries, this IP (Intellectual Property) -based intelligence has gained increasing awareness over the past decades (Rivette & Kline, 2000). As technologies have become more and more complicated over the years, inventions nowadays are often described by a vast amount (sometimes thousands) of patents (Kur & Dreier, 2013). In order to convert this ever-growing database of patents into actionable intelligence, automated technologies for patent analysis have been developed. Furthermore, also tools capable of text mining for technology mapping came to the market. The output of these mapping tools represent the technological content of a certain set of patents in a map that borrows its appearance from cartography (Trippe, 2015). By doing this, huge amounts of patent records (up to millions) can be analysed, clustered based on their technological content and represented in an easy to interpret manner. This all can be done in a fraction of the time it would take to do this manually. The question raised however, what the quality of automated patent mapping is and how can it be used as a source of competitive or business intelligence?

An answer to this question was found during a case study at Philips, executing research into two tools capable of automated patent mapping: ThemeScape and Orbit. Due to their installed base, these two tools where assumed to be indicative for the state of the art in patent analysis. These tools where assessed for their quality by a measure of usability (Bevan, 1995). Applications for patent maps were found in the support of clustering in IP analyses, fast due diligence, scoping of R&D projects and exploration for Merger and Acquisitions (M&A). The usability was then quantified by measures of effectiveness, which entailed an accuracy and efficiency assessment by means of an experiment. The results showed that the tools where highly efficient but seemed to lack accuracy in their technology clustering. Since the definition of a technology is subject to change (Wittgenstein, 1953) and a subjective matter, measuring the quality of automated clustering of technologies is a challenging task.

Lastly, a research was executed into the technologies and design choices within the software of the tools. These choices appeared to be an explanation for the inconsistent character of the output of the tools, that makes them less applicable in the foreseen applications. There is also room for improvement in the technologies behind the tools.

Concluding, an answer to the research question could be formulated as follows: Due to the technical limitations, inconsistency in mapping and fluctuating performance of accuracy, automated patent mapping tools can only be seen as effective in the support of the creation of competitive / business intelligence rather than being a source of it itself. The mapping output is too susceptible to inconsistency for being a genuine source of IP-intelligence, the same holds for the clustering performance.

**Key words:** IP intelligence, competitive intelligence, business intelligence, automated patent mapping, patent content analysis, big (IP) data visualization, text mining, dimensionality reduction
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>Business Intelligence</td>
</tr>
<tr>
<td>CASA</td>
<td>Centre for Advanced Spatial Analysis</td>
</tr>
<tr>
<td>CI</td>
<td>Competitive Intelligence</td>
</tr>
<tr>
<td>CPC</td>
<td>Cooperative Patent Classification</td>
</tr>
<tr>
<td>CT</td>
<td>Computer Tomography</td>
</tr>
<tr>
<td>DI</td>
<td>Derwent Innovation</td>
</tr>
<tr>
<td>DR</td>
<td>Dimensionality Reduction</td>
</tr>
<tr>
<td>DWPI</td>
<td>Derwent World Patent Index</td>
</tr>
<tr>
<td>EBITA</td>
<td>Earnings Before deduction of Interest, Tax and Amortisation</td>
</tr>
<tr>
<td>ECLA</td>
<td>European Classification</td>
</tr>
<tr>
<td>EPC</td>
<td>European Patent Convention</td>
</tr>
<tr>
<td>EPO</td>
<td>European Patent Office</td>
</tr>
<tr>
<td>FDGD</td>
<td>Force-directed Graph Drawing</td>
</tr>
<tr>
<td>IP</td>
<td>Intellectual Property</td>
</tr>
<tr>
<td>IP&amp;S</td>
<td>Intellectual Property &amp; Standards department</td>
</tr>
<tr>
<td>IPC</td>
<td>International Patent Classification</td>
</tr>
<tr>
<td>IPgen</td>
<td>IP-generation-project</td>
</tr>
<tr>
<td>IPRs</td>
<td>Intellectual Property Rights</td>
</tr>
<tr>
<td>M&amp;A</td>
<td>Merger and Acquisition</td>
</tr>
<tr>
<td>MDS</td>
<td>Multidimensional Scaling</td>
</tr>
<tr>
<td>NMDS</td>
<td>Non-metric Multidimensional Scaling</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PCT</td>
<td>Paris Convention Treaty</td>
</tr>
<tr>
<td>PNNL</td>
<td>Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>RBV</td>
<td>Resource Based View</td>
</tr>
<tr>
<td>SNE</td>
<td>Stochastic Neighbour Embedding</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for the Social Sciences</td>
</tr>
<tr>
<td>TRIPS</td>
<td>Agreement on Trade Related aspects of Intellectual Property Rights</td>
</tr>
<tr>
<td>USPTO</td>
<td>United States Patent and Trademark Office</td>
</tr>
<tr>
<td>WIPO</td>
<td>World Intellectual Property Organization</td>
</tr>
<tr>
<td>WTO</td>
<td>World Trade Organisation</td>
</tr>
</tbody>
</table>
Contents
Acknowledgements .................................................................................................................. I
Summary ................................................................................................................................... II
Abbreviations .............................................................................................................................. III
1. Introduction .............................................................................................................................. 6
   1.1. Motivation and research ................................................................................................. 6
   1.2. Research objective .......................................................................................................... 9
   1.3. Structure ......................................................................................................................... 12
2. Literature review .................................................................................................................... 13
   2.1 IP law................................................................................................................................ 13
   2.2 Philips IP policy ................................................................................................................ 17
   2.3 The complex world of patents ......................................................................................... 18
   2.4 Towards (efficient) IP-intelligence by automation ........................................................... 20
   2.5 Technologies of automated patent mapping .................................................................... 28
      2.5.1 Building the vector model ....................................................................................... 29
      2.5.2 Clustering ................................................................................................................. 30
      2.5.3 Dimensionality reduction ....................................................................................... 30
      2.5.4 Mapping .................................................................................................................... 33
3. Methodology .......................................................................................................................... 34
   3.1 Research strategy ............................................................................................................. 34
   3.2 Case description ............................................................................................................... 35
      3.2.1 Patent analysis process at Philips ............................................................................ 36
      3.2.2 Patent mapping tools at Philips .............................................................................. 38
   3.3 Research method and data gathering ............................................................................... 39
      3.3.1 Capabilities .............................................................................................................. 41
      3.3.2 Effectiveness ............................................................................................................ 42
      3.3.3 Introspection ............................................................................................................ 49
4. Research and Results ............................................................................................................. 56
   4.1 Capabilities ....................................................................................................................... 56
      4.1.1 Functionalities ......................................................................................................... 57
      4.1.2 ThemeScape .......................................................................................................... 58
      4.1.3 Orbit ...................................................................................................................... 62
      4.1.4 ThemeScape and Orbit ......................................................................................... 64
      4.1.5 Applications ............................................................................................................ 67
   4.2 Effectiveness ..................................................................................................................... 74
      4.2.1 Results .................................................................................................................... 74
1. Introduction

1.1. Motivation and research

Over the years, the value of intelligence for corporations is being recognized (Miller, 2010); in this case intelligence refers to an actionable, refined form of data (McGonagle & Vella, 2012). For example, the ability to analyse market developments instead of merely react to them can have major implications for businesses (Miller, 2010). In 1979, Porter introduced the world to his ‘Five Forces Analysis’, which was designed for analysing competition in order to develop a strategy for gaining long-term profits (M. E. Porter, 2008). With his evolutionary model, Porter emphasized on business risks that can come from the competitive environment in which a certain company is operating (Pargaonkar, 2016). Analysing the competitive environment can result into valuable intelligence that can be used in order to reduce business risks. Porter’s model turned out to be the basis for the extensive literature base on competitive intelligence for companies existing nowadays. “Competitive Intelligence (CI) involves the use of public sources to develop data on competition, competitors, and the market environment. It then transforms, by analysis, that data into intelligence” (McGonagle & Vella, 2012, p. 9).

Rodenberg (2007) adds to the definition of CI that it should be actionable, future oriented and that it is focussed on the key drivers of change in the market. These drivers of change and future orientation relate back to Porter’s five forces model, monitoring these five forces (competitors, new entrants, substitute products or services, buyers and suppliers) should prevent companies to be surprised by changes in the external environment (Rodenberg, 2007). Nowadays, in rapidly changing markets, the responsiveness of a company to external change appears to be key for maintaining its position (Christensen, 1997; Downes & Nunes, 2013; O’Reilily & Tushman, 2004). “Competitive intelligence is never finished for the simple reason that the outside world is constantly changing” (Rodenberg, 2007, p. 27). With actionable intelligence is meant that the intelligence can be used as a support or bases for decision making (Rodenberg, 2007). In other words, the intelligence should be useable in a business setting. A good way to make intelligence actionable is by graphical visualisation. “Graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers [data] is to look at pictures of those numbers” (Tufte, 2001, p. 9).

The value of competitive intelligence can also be seen by looking at it from a Resource Based View (RBV), as this was described by Barney (1991). This view argues that the
performance of firms depends on the resources they possess. A firm’s resources can be defined as “all assets, capabilities, organizational processes, firm attributes, information, knowledge etc. controlled by a firm that enable the firm to conceive of and implement strategies that improve its efficiency and effectiveness” (Barney, 1991, p. 101). By the capability of analysing its competitive position, a firm can generate information and knowledge, which it can use to refine its competitive strategy, in order to strengthen the future position of the firm in the market. CI is a firm’s resource and in that way, according to RBV, it can contribute to the firm’s performance as a source of competitive advantage.

By some authors, competitive intelligence is mentioned in the same breath as business intelligence (BI) as they attach the same (McGonagle & Vella, 2012) or very interrelated meaning to both terms (Negash, 2004; Pargaonkar, 2016). In many other literature sources CI is seen as a specialized field of BI, that focuses mainly on the external competitive environment. Ranjan (2009) provides an overview of authors that give a definition that copes with this last view. In this research the difference between the two will be acknowledged and CI is treated as this specialized form of BI. In this research the definition by Gangadharan and Swami (2004, p. 140) will be adapted; “BI is a term that encompasses a broad range of analytical software and solutions for gathering, consolidating, analysing and providing access to information in a way that is supposed to let an enterprise’s users make better business decisions”. This research will cope with patent-based business and competitive intelligence, with a strong focus on its competitive value. The competitive value of patents is very well indicated in the book by Rivette and Kline (2000) by many real-life cases.

Over the years, the data sources to build CI upon have developed and changed, together with methodologies to analyse this data. In the recent years, awareness on the potential for intellectual property (IP) to be a source for competitive intelligence has raised. Mainly because in many industries knowledge has become the most important resource for the execution of the business (Gračanin, Kalac, & Jovanović, 2015; Newell, Robertson, Scarbrough, & Swann, 2009). This change in focus has also been recognized by Rivette and Kline (2000) who, with their book, created greater awareness of the importance of patents in the corporate world. As “the competitive battles once fought for control of markets and raw materials are now increasingly being waged over the exclusive rights to new ideas and invention” (Rivette & Kline, 2000, p. 2). In a business-world where knowledge can majorly influence a company’s market position, IP (in particular patents) have become of great importance. The area of IP

---

information operates on the cross-section of legal, scientific-technical and economic aspects (Stock & Stock, 2006). “As part of the ‘patent deal’ between inventors and society, inventors have to disclose information about their inventions in exchange for patent rights” (Granstrand, 1999, p. 289). Patent databases open up information about, for example, inventions, inventors, owners, inventions that relate to each other, moment of invention, legal status, and many more. Therefore, patents offer a great source for CI and BI, that should not be neglected by any company. Some examples of how this information can be used are: avoiding overinvestment in R&D, coordination of inventive labour, (Granstrand, 1999), finding collaborating partners, showing technology trends (Granstrand, 1999; Yang, Akers, Yang, Klose, & Pavlek, 2010), showing leading patent assignees (Yang et al., 2010), supporting merger and acquisition decisions and supporting high-level policy matters (Trippe, 2015). Large patent databases are accessible from all over the world and hold great potential to support decision making for businesses and government (Trippe, 2015). What remains is the question of how to retrieve useful information from the world-wide patent database (Gračanin et al., 2015)?

An answer to this question can be found in the field of data processing. “Patent data needs to be converted into effective IP, business and competitive intelligence in order to be used in the corporate world” (Pargaonkar, 2016, p. 21). As mentioned before, intelligence should be actionable (Rodenberg, 2007). Converting data into intelligence is done by analysing data, or as Ackoff (1989) phrased it in the article that formed the basis for his famous ‘wisdom pyramid’; “Information [intelligence] consists of processed [analysed] data, the processing directed at increasing its usefulness” (Ackoff, 1989, p. 1). The analysis of patents falls into the field of what Porter and Cunningham (2005) introduced as tech mining. “Tech mining is about information [scientific and technological information] to see patterns, detect associations, and foresee opportunities. The derived knowledge can help make better plans, designs, and decisions, thereby gaining significant competitive advantage”. A specific field of tech mining is to apply analysis techniques to patent databases. To make these analyses actionable, patent analysis reports are created that form a source of CI and/or BI, or simply, IP-based intelligence.

Patent analysis reports typically contain visual representations of the information that was gathered with a state-of-the-art search in the form of graphics and charts supported with textual information about these visualisations (Yang et al., 2010). As over the years the amount of patents and the complexity of their content increased, also the means for analysing big chunks of patent data developed. Due to the complexity and volatility of innovation-based environments, efficient and effective means for analysing patents are desirable. Some of the most advanced technologies out there are tools that are capable of analysing patent content in
an automated manner in order to ‘map’ patents. “Patent mapping is essentially the visualization of the results of statistical analysis and text mining processes applied to patent documents. Patent mapping allows you to create a visual representation of information from and about patent documents in a way that is easy to understand” (EPO, n.d.-d). In general, automated tools for patent mapping are relative fast, their output is easy to comprehend and the user should be able to generate new insights based on these patent maps. Thus, these tools hold the potential to contribute to a company’s IP based intelligence. However, because certain tools are highly automated, it is hard for the user to validate the quality of the output. To put it differently, the user of a patent mapping tool can upload a set of patent documents, the tool analyses these documents by running its algorithms and when finished, the user retrieves an output. However, what exactly happens in the tools in order to create the output is vaguely described or simply unknown, which raises the question whether the output does indeed show what the user is looking for.

Research will therefore be conducted into the true value of tools capable of automated patent mapping in the creation of IP-based intelligence. This will be done during a case study at the Intellectual Property & Standards-department (IP&S) within Philips. Philips is a multinational company in providing health care systems and the production of consumer electronics. By the courtesy of Philips, licenses were provided that enabled the use of ThemeScape (as part of the Derwent Innovation patent analysis module) and the analysis module of Orbit Intelligence for this research. Both of these tools can be used for automated patent mapping. “A patent map generally represents a graphical representation of a data collection that borrows characteristics from cartography. […] A map paradigm is used to represent similarity between documents or concepts since the human mind is used to and can readily understand the use of maps to correlate distance between two items” (Trippe, 2015, p. 29). IP&S handles all IP matters related to Philips businesses’ (Philips, n.d.) and has an extensive experience in the generation of IP based intelligence, provided by IP-analysis in the form of (client-specific) IP-analysis-reports. Currently, patent maps are not used in these reports, because for them (and many other companies) the question remains what can be learned from automated patent maps and whether these maps indeed show valuable information.

1.2. Research objective
The worldwide patent database is expanding continuously and the complexity of the technologies described in patents increases. Having access to artificial brains for reading through vast amounts of patents to show mutual relatedness and make their content more
comprehendible in a very short time is an exciting development. Especially for innovative companies that recognise the value of patents for building intelligence. The artificial brains referred to are software tools capable of automated patent mapping. However, as exciting as these technologies may sound, the true brightness of these ‘brains’ is not yet assessed. The body of literature on the applications of automated patent mapping in business (workflows) is limited in size and structurally lacks rigorous validation of the tools that create these maps (see Chapter 2). Former research merely focuses on the potential user groups and the apparent strengths and limitations of automated patent mapping. This thesis research builds on the opinion that only by a rigorous validation of the quality, the true added value of automated technologies for companies can be assessed.

This thesis research furthermore, builds on the need for deeper understanding of automated patent mapping applications in IP-analysis-processes for the generation of IP-based intelligence. Automated mapping can be seen as a means for time reduction in the generation of IP-intelligence for policy and decision-making. Besides this, automated mapping holds the potential for generating new insights into large sets of patent data. A set that would normally be too big for people to comprehend can be visualized in such a way that it almost speaks for itself in how to interpret the content and relationships.

The purpose of this research is to investigate and to gain insight into the quality of automated, patent based, technology mapping. Based on this assessment, a judgement will be made on to what extent it can function as valuable resource of intelligence. In other words, automated patent based technology mapping will be validated on being a source of IP-intelligence. Figure 1 provides a schematic representation of where the need for this research emerged from.
Figure 1: There are many ways to convert patent data into CI, in this research automated patent mapping as a means for this conversion will be validated.

The research objective, as it is described above, leads to the following research question:

**What is the quality of automated patent mapping and how can it be used as a source of competitive or business intelligence?**

The very broad scope of this research question is narrowed by focussing the research on two tools capable of automated patent mapping, which were found descriptive for the state of art in this field of technology. Furthermore, the scope is narrowed by researching this question at one company: Philips. Philips was found to be a good representation for highly innovative companies in which the value of IP-based competitive and business intelligence is recognized, because of the extensive experience Philips holds in IP-analytics. Also, Philips was one of the pioneering companies in this area. In the methodology chapter the motivations and rationales for the research, that was designed for answering this question, are extensively elaborated upon.

With answering this research question a better understanding of the applicability of automated patent mapping in business will be provided. To formulate a comprehensive answer to this research question, several topics will be investigated with different research methods. This leads to methods triangulation, which can be described as using multiple data sources and collection methods to generate an in-depth understanding of the phenomena under investigation. Topics under investigation will be the *quality* assessed by: *capabilities* and *effectiveness* of the tools and *techniques* the tools comprehend. Furthermore, some real life
cases will be described in which patent maps could be a resource for decision-making actors within a company. The combination of outcomes of the researches above will provide a comprehensive basis on which an answer to the research question can be formulated.

1.3. Structure

In the next chapter an introduction is given to the patent system, laws and treaties, guided by Philips’ own IP related (legal) processes. This in order to provide the reader with sufficient understanding of intellectual property matters. Next, the literature review will be extended by providing all relevant literature on IP analysis towards intelligence and the technologies in the field of automated patent mapping. In the third chapter the research methodology will be described and in chapter four the executed researches will be explained. Chapter four starts with an extensive elaboration on the functionalities in the tools, followed by real-life applications for the tools found in the literature and in Philips internally. The chapter than continues towards methods of validating the outcome of the tools, in order to verify its applicability in the described cases. Altogether, this will lead to the fifth concluding chapter, in which the conclusion from the research is described and the managerial implications of this conclusion are elaborated upon. This research ends with a discussion.
2. Literature review

In this chapter a literature review will be presented that comprises all relevant literature on IP analysis for technology mapping. This review gives a state of art literature overview of those concepts that are essential for answering the research question. Also, indications about where the findings in the literature are insufficient, for answering the research question, are given. This is where this research finds its major academic relevance.

First, an introduction into the global patent systems will be given in paragraph 2.1 to 2.3, guided by Philips’ own processes and policies on patents. This introduction into the global patent system is most relevant to readers that are relatively new to this field, in order to understand where the research is about. Besides, concepts will be introduced and extensively explained, such that the reader is informed about their function and their potential impact on the research. This introduction will also show where the need for automated analysing technologies emerges as it describes how complex the field is. In paragraph 2.4, an introduction is given into automated technologies in patent analysis. This paragraph emphasizes on scholars that performed research into tools for automated patent mapping. Focus of these researches is on the users applications of the tools and their perceived capabilities. However, a gap in the literature was found related to the validation of these perceived capabilities. In many literature sources, a lack of scepticism against the true value of the tools’ visualisations was found and also real-life user cases were absent, only speculative ones were found. True validation is lacking in general. In paragraph 2.5, a high-level technical overview of automated mapping tools is provided. As the techniques and design choices in the tools lead to how the data (patent records) is presented to the user, insight in these techniques is necessary for making a comprehensive quality judgement of the tools, in order to validate them.

2.1 IP law

Patents are applicable to inventions, however only to those inventions that are in the technological field, are new, involve an inventive step and are applicable of industrial application (WTO, n.d.). The word patent derives from the Latin word ‘patere’, translated as: lying open. This because the disclosure of the invention is a mandatory condition for protection. In return to the disclosure the inventor receives an exclusive right for a limited time period (Kur & Dreier, 2013). A patent can be defined as “a legal title granting its holder the right - in a particular country and for a certain period of time – to prevent third parties from exploiting and
invention for commercial purposes without authorization” (EPO, 2017). Strictly speaking, a patent is not an invention or a technical document but a legal right with possible economic value. A patent does not by definition give the holder the right to sell or manufacture the invention, but the right to exclude others from doing that (Granstrand, 1999). In article 27 and 28 of the TRIPS agreement² is described to what patents are applicable and which rights a patent provides to its holder:

“Patents shall be available for any inventions, whether products or processes, in all fields of technology, provided that they are new, involve an inventive step and are capable of industrial application.

1. A patent shall confer on its owner the following exclusive rights:
   - where the subject matter of a patent is a product, to prevent third parties not having the owner’s consent from the acts of: making, using, offering for sale, selling, or importing for these purposes that product.
   - where the subject matter of a patent is a process, to prevent third parties not having the owner’s consent from the acts of: making, using, offering for sale, selling, or importing for these purposes at least the product obtained directly by that process.

2. Patent owners shall also have the right to assign, or transfer by succession, the patent and to conclude licensing contracts” (WTO, n.d.).

Patents are a particular form of intellectual property rights (IPRs). IPRs protect immaterial goods, which means goods that are intangible or non-physical (Kur & Dreier, 2013). Kur and Dreier define immaterial goods as goods “that are mostly the product of a creative mental human activity in the industrial, scientific, literary and artistic field” (2013, p. 2). Next to patents, intellectual property rights cover also: trademarks, trade secrets, copyrights, design rights and artistic rights³ (Granstrand, 1999). In a corporate asset structure, patents categorize under the heading of ‘intangible assets’. The management of these intangible assets relates to ‘knowledge management’, as knowledge (generated by human capital, innovations or acquisition of IPRs) can generate the firm’s IPRs (Scheffer & Zieger, 2005). Valuing these intangible assets can be done by the application of expert valuation techniques, such as for example calculating the net present value (Berk & DeMarzo, 2013).

² Agreement on Trade Related aspects of Intellectual Property Rights
³ These other categories of IPRs will however be outside the scope of this research and therefore not elaborated further upon. Interested readers could delve into these topics by reading for example the book by Kur and Dreier (2013), that gives a comprehensive overview of general IPRs in Europe.
The data used in this research will mainly originate from patents, therefore next the process of filing, patent law and the characteristics of patents will be described. For this, mainly the legal definitions and descriptions used by global organizations and agreements will be investigated as they are most similar to the national and regional definitions, in order to prevent loosing ourselves with too much exceptions and legal details. Besides that, they cover the situations with which Philips is dealing, as it is operating globally.

In general, three layers of patent laws and procedures can be distinguished (Cook, 2015). The top layer, or the top of the hierarchy, consist of international treaties. These treaties are the Paris Convention Treaty (PCT), as it was administered by the World Intellectual Property Organization (WIPO), in 1970 (Cook, 2015). As a specialized agency of the United Nations the WIPO has the task to promote the protection of intellectual property globally (Kur & Dreier, 2013, p. 17). The PTC harmonized the application procedure for patents, by setting up a central filing procedure and an international search report for prior art, after which the application can proceed in national offices (Cook, 2015; Kur & Dreier, 2013). An international search report for prior art identifies patent documents and technical literature that might influence the (scope of) patentability of the invention. The report also consists of documents that have relevance to critical patentability questions about its novelty and inventive step (see above mentioned article 27 and 28 if the TRIPS agreement) (WIPO, 2017). Documents that question an invention’s novelty are also referred to as ‘prior art’. Prior art is any evidence that an invention is already known (EPO, n.d.-f). “An invention is considered to be new if it is not part of the state of the art. The state of the art shall be held to comprise everything made available to the public…before the date of filing” (EPO, n.d.-a). So, anything known to the public domain before the filing date of the patent can be referred to as prior art and can influence the patentability of the invention. To summarize, the search report evaluates the chances of obtaining a patent for the invention (WIPO, 2017). The WIPO covered, besides procedural issues, also agreements to not discriminate against other countries in the granting of patents (Cook, 2015). More critical, this forced countries “to treat patent applications filed in one contracting country as having been filed on the same date in a second contracting country”. This concept is also referred to as the ‘priority date’ or ‘right to priority’, or just: priority. To put it differently, the right to priority enables an applicant to file a subsequent application for the same invention (within a certain period) in a different country and obtain the same date of filing of the earlier (first) application for this subsequent filing. After singing the PCT, more international harmonization of patent law occurred by singing of the agreement on TRIPS in 1994 by the World Trade Organization (WTO) (Cook, 2015). The TRIPS agreement did set
new and higher minimum standards of law for patents (for example the above cited article 27 and 28 if the TRIPS agreement). These standards are applicable for all the 144 members of the WTO that signed the TRIPS agreement.

To summarize the above, in the international layer the filing procedure is mainly prescribed in the PCT by the WIPO. The law on patents was globally more harmonized between the members of the WTO via the TRIPS agreement. The centralizing effect of the PCT is found in the centralization of applications, an international search report, and a preliminary examination (on request of the applicant) (Kur & Dreier, 2013, p.91).

The second layer consists of the regional treaties, for example the European Patent Convention of 1973. This convention constructed the basis for more uniform patent law and systems throughout Europe. It established the possibility to prosecute patents via the European Patent Office (EPO) in order to secure a set of national patents (Cook, 2015). The EPO was founded during the European Patent Convention (EPC) (Kur & Dreier, 2013, p. 65) and functions as a system for obtaining patent protection in Europe in an efficient way. The EPO offers a common granting procedure for all members (which also include a number of non-EU states), which leaves national patent laws unaffected (Kur & Dreier, 2013, p. 87). Via the EPO, the inventor can receive a granted ‘European patent’ based upon a single application, examination and registration procedure. With a granted European patent one can let the patent validated in each EPC-country designated by the patent holder. National validation of the patent has some certain formal requirements, such as translating the patent, in each individual country where protection is sought (Kur & Dreier, 2013, p. 88).

To summarize, “in its centralizing effect, the EPC goes beyond the PCT system, whereas the PCT only provides for a centralization of applications, an international search report, and, on request of the applicant, a preliminary examination, the EPC unifies the entire examination process and results in the grant of a common title, the ‘European Patent’” (Kur & Dreier, 2013, p. 91).

The third layer is the national layer, which consist of each countries’ individual patent act. “The European patent once granted will subsequently enter into its national phase and live on in the form of one or several separate national patent(s)” (Kur & Dreier, 2013, p. 104). Before these international treaties came into force an actor seeking patent protection in several countries had to file separate applications in each and every country, which implies high costs for multiple representations by patent attorneys in all the different countries (Kur & Dreier, 2013, p. 90). These different layers offer efficient routes via which a company like Philips can file its patent applications in order to obtain protection in countries all over the world.
The following paragraphs will shown how these layers in patent law have contributed to the complexity of the whole system, as terms like ‘patent family’ and ‘classification codes’ will be introduced to the reader. The next paragraph, however, will first disclose how Philips copes with this international system.

2.2 Philips IP policy

After an invention has occurred at Philips the first step for the inventor(s) in order to receive patent protection is to disclose the invention to an Intellectual Property Council (IP-council) at IP&S. The IP-council will evaluate the invention and has four actions it can take. The IP-council will put the invention disclosure in Philips’ Archive when the IP-council does not see any potential for the invention. The IP-council can choose to Publish the invention, after which it becomes part of the public domain. This can be done when the invention has not much potential for Philips’ current businesses and Philips wants to prevent others to be able to protect this invention in the future. Since, the invention becomes part of the public domain, nobody will be able to protect the invention because the disclosure counts as prior art. The third option the IP-council has is a Novelty Search request, when the (complete) novelty of the invention is not yet trusted. Similar to the earlier mentioned ‘search report’, a novelty search is mainly used to find documents that are referred to as ‘prior art’. Supported by a novelty search the IP-council can then again judge whether the invention is new and available for patent protection. If so, the IP-council can pick the fourth option, and go over to File the invention at a patent office.

In general, a company can file nationally, regionally or internationally. When filing internationally, a company can choose to file via the EPC (Europe) or the PCT. Philips usually files internationally via the PCT route. A visual representation of a PCT filing process example is shown in Figure 2. The country codes are in the drawing by randomness and do not tell anything about Philips’ filing strategy. After 18 months from filing, the PCT will publish the content of the application and the invention then is disclosed to the world as it becomes part of an open accessible database (WIPO, 2017). This publication will acts as prior art against applications of similar inventions in the future (EPO, n.d.-e).
Within 30/31 months after filing, the patent enters its national phase based on its PCT priority (The black square visualises the priority date in Figure 2). Very strictly, all the patents that emerge out of one priority are referred to as a patent family, or *family*. Put differently, one may say that a family comes forth out of the same invention. However, the next paragraph will show that something as straightforward as a family definition, leaves a lot of room for ambiguity in practice. The complexity of the patent system will be highlighted by the existence of different means to group or classify patents, which largely emerged from the earlier described layers in patent law.

### 2.3 The complex world of patents

Patent structures have become more complicated over the years, since “in the early days, new machines or tools where basically covered by a single patent, where nowadays more than a thousand patents may apply to a single product” (Kur & Dreier, 2013, p. 84). Literature and authorities are therefore not always uniform in defining a family, in practice the definition of a family is not so straightforward. The EPO defines a patent family as “a collection of patent applications covering the same or similar technical content” (EPO, n.d.-c), this particular definition already leaves room for debate as similarity can be subjective. The WIPO (n.d., sec. P) defines a patent family as “a collection of published patent documents relating to the same invention, or to several inventions sharing a common aspect, that are published at different times in the same country or published in different countries or regions. Each patent document in such a collection is normally based on the data for the application(s) on which the basis for its priority right has been claimed”. The WIPO provides a very broad definition of patent
families in its definition. The lack of rigor in defining patent families becomes more obvious in so-called ‘complex structures’ of a patent application (Espacenet, 2017). These complex structures occur out of patents with a different scope of protection but still are sharing at least one priority. When patent applications are filed in different countries, various earlier applications can be cited as priorities. On top of that different claims may get rejected or accepted by the different offices during the granting period (Espacenet, 2017). This indistinctness in family definition is a strong indication of the complexity of the world of patents in general and the challenge this brings for adding structure. Besides the above mentioned family definition, every tool under investigation uses its own family definition.

In terms of complexity, more or less the same problems hold for the different international classification systems for patents documents. As it is often easy to justify a single patent to be grouped into more than one class. “The purpose of a patent classification system is to organize patent collections according to their technical application, structural features, intended use or the resulted product produced by a process” (Meireles, Ferraro, & Geva, 2016, p. 6). Generally, patent classifications occur in hierarchical structures of patent documents where the detail of description increases by going deeper into the hierarchy. Some examples of such classifications are the International Patent Classification (IPC), maintained by the WIPO, the European Classification (ECLA), maintained by the EPO. “While the International patent classification seems to be more oriented toward the [uniform] publication of patents, the European patent classification is more focused on [uniformly] supporting patent information search in the context of a patent application” (Meireles et al., 2016, p. 8). A joint partnership between the United States Patent and Trademark Office (USPTO) and the EPO resulted in the Cooperative Patent Classification (CPC). As “each of the classification systems was developed with a different underlying philosophy” (Meireles et al., 2016, p. 9). There are thus, more than one way to cluster and classify patent documents, depending on what the classification was intended for. Which is not very different for clustering patent documents with automated software. The tools under investigation mainly claim to provide an overview of technologies described in a certain set of patents (Clarivate, 2018; Questel, n.d.-b). As defining an invention (in the family definition), and classify records (with classification codes) appeared to be not a straightforward task, the very same holds for technology clustering based on patent text. Mainly two reasons are identified that cause this complexity; the character of patent text and the definition of Technology.

As described before, patents are a legal means, or a legal right with certain economic value, rather than a document describing a technology. However, the latter is also included in
patent documents, there is a difference in the character of the technology description with other
documents disclosing a technology, such as academic publications. Where academic
publications in general attempt to be as clear as possible, “it is known that patents are difficult
to read and comprehend, not only because they usually contain highly specialized technical
descriptions, but also because legal terms are used to define the scope of the invention”
(Meireles et al., 2016, p. 4). “As long as authors can explain the invention in terms that the
examiner can understand and they apply the same language consistently throughout the
document, they are within the boundaries of the law” (Trippe, 2001, p. 62). Hence, patent
documents are very complex and reading or mining a patent collection is a big challenge
(Meireles et al., 2016).

As the goal of the tools under investigation is to map out and cluster technologies, the
definition of a certain technology should be clear. Also, to make a judgement about the quality
of the tools’ performance. However, this definition is generally filled with ambiguity. This
ambiguity finds its origin in the meaning of words. Certain words can describe technologies,
however the interpretation of words and the extension of the meaning attached to it is a
subjective matter and a subject to change. This changing character of the meaning of words is
invigorated by findings philosophic literature. Wittgenstein explained the meaning of a word
as follows; “the meaning of a word is its use in language” (1953, p. 19). As the use of a word
in language can change over time, meaning does as well. Further, the subjectiveness of the
meaning awarded to a word can be indicated by its connotation. “The connotative meaning of
a word includes the feelings and ideas that people may connect with that word” (Cambridge
Dictionary, n.d.-a). These ideas are mainly built on past experiences with the word. So,
especially in the field of complex technology definitions, people’s meaning attached to certain
technologies might differ and can still be ‘right’. In the next paragraph, will be elaborated upon
advanced IP analyses that are applied to make sense out of the above described complex field
of patents, in order to generate effective IP intelligence for business purposes.

2.4 Towards (efficient) IP-intelligence by automation

In 1999, Granstrand mentioned that “new methods for collecting, analysing and presenting data
and information are being developed. Computers and telecommunications will expand the
availability of empirical information and its amenability to analysis” (Granstrand, 1999, p. 290).
This prediction appeared to be right for patent data as we next monitor the development in
automated analysis techniques. This monitoring also comprises those publications that are on
the assessment of automated techniques for patent analysis. In these academic publications, a literature gap was found that will be filled by this thesis research: a rigorous validation of automated tools before making an assessment of its usability in practice.

In 2003, Trippe published an overview of common patent analysing tasks together with an description of tools that are likely to be useful in performing these tasks (Trippe, 2003). These tasks comprise analysing activities related to: data clean-up, list generation (count of patent related metrics), clustering of structured data (meta-data), clustering of unstructured data (such as text fields), mapping of document clusters and the related functionalities within these maps and discernment of descriptive wording in patent text. Dou, Leveillé, Manullang, & Dou Jr (2005) created a list of patent analysis tools with a focus on analysing bibliographic (meta-) data. Some years later, Yang, Akers, Klose, & Barcelon Yang (2008) presented an overview of some key text mining and visualization tools, containing a description of their perceived strengths, potential limitations and suggestions about user groups that may benefit from the tools. In 2015, Trippe published a report containing guidelines about the preparation of landscaping reports (or, IP-analysis-reports), this publication by Trippe (2015), contains an updated overview of analysis tool providers (Trippe, 2015). What is lacking in these studies on automated software for IP-analysis, is a rigorous assessment of their quality when looking into the potential business applications of tools.

Over time, “there appeared to have been a shift in focus towards semantic analysis algorithms, where the focus used to be on statistical analysis techniques” (Yang et al., 2008, p. 280). Trippe (2015) referred to statistical analyses in patent landscape reports as mainly “counting items in certain patent information fields” (A. Trippe, 2015, p. 92). Statistical patent analysis are mainly executed on the bibliographic (meta-) data within the patent documents (e.g. document identification data, priority date, publication data, classification). Many of these bibliographic data is readily applicable for statistical analysis, since it is presented in structured fields (Trippe, 2015). The quality of the statistical analysis and its visualization (e.g. graphs and charts) depends on the preparation of the data before the analysis (Trippe, 2015; van der Ligt, 2017). Semantic analysis are a certain approach that can be utilized when one is applying text mining techniques (Abbas, Zhang, & Khan, 2014). In data mining (text is a variant of data), “mining implies extracting precious nuggets of ore from otherwise worthless rock” (Hearst, 1999, p. 3). According to Hearst (1999), especially in the case of text mining this metaphor can be taken seriously. As, “text mining is often regarded as a process to find, implicit, previously unknown, and potentially useful patterns from a large text repository […] after supplemented with additional information and interpreted by experienced experts, these patterns become
important for decision-making” (Tseng, Lin, & Lin, 2007, p. 5). This is the same for text mining of patent texts. The purpose of text mining tools for patent mapping is to find shared patterns in patent content as a basis for clustering of the patent records. These clusters will then be represented in a map and used for gaining different insights as a contribution for decision-making processes.

Yang et al. (2008) grouped the different text mining and visualization tools for patents into three categories based on their text mining capabilities; tools capable of mining unstructured (e.g. full-text patent documents, news or journal articles), tools capable of mining structured data (e.g. patent front pages) and tools capable of mining hybrid data (e.g. full patent documents). The tools under investigation in this research can be classified in the second group. In many of the above mentioned academic publications, several tools are evaluated, mainly on the bases of its capabilities presented by its vendor. In this research the assessment of these tools will be taken one step further, by assessing how the tools perform in a real life setting (by making the performance measurable).

Table 1 provides a list of all the tools similar to the tools under investigation in this research. The table only shows tools that are capable of text mining for the purpose of mapping patent content and/or visualizations of technology clusters. Whenever the reader is interested in a broader overview of patent analysis tool the WIPO-report; ‘Guidelines for preparing patent landscape reports’ by (A. Trippe, 2015, p. 105 - 108) provides an extensive list of the state-of-the-art of analysis tools. ThemeScape and Orbit capture technologies from patent content and present them in a topographical visualization, in which each patent document is represented as a point in a landscape. The proximity of the points towards each other provide an indication to what extend the documents, and thus the technologies described in them, are related to each other. Groups of points, or clusters (multiple records describing highly related technologies) are labelled with words indicative for the technological field described in them. Islands with hills and sea areas represent respectively high and low number of documents related to a particular subject. With the functionalities offered in the tools the user can add a dimension to the map, such as highlighting certain groups of points as they match a certain condition (e.g. show in which technology fields competitors are active or visualize the legal status of the patent documents in certain technology fields). Below, in Figure 3 and 4, screenshots are provided of the output maps of both tools to provide an indication of such maps actually look. These maps are based on a patent dataset retrieved from the database in both tools that comprise patents that are alive and have Philips as assignee.
Figure 3: Patent map related of Philips’ active patents, using the ThemeScape software module
Figure 4: Patent map related of Philips’ active patents, using the Orbit software module
Clarivate Analytics and Questel, the vendors of respectively ThemeScape and Orbit have both an extensive experience in the handling of large patent data. Having access to these tools offers the possibility to research whether automated patent mapping can by any means contribute to a company’s IP-based intelligence. To phrase this differently; research needs to show whether these maps indeed disclose reliable and valuable information from big patent data in a fast and easy to comprehend manner, or might the user as well read its tealeaves to find the same quality of information?
<table>
<thead>
<tr>
<th><strong>Vendor</strong></th>
<th><strong>Tool</strong></th>
<th><strong>Description</strong></th>
<th><strong>Link</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarivate Analytics</td>
<td>Derwent Innovation</td>
<td>“Integrated patent and non-patent literature searching and analysis system, featuring ThemeScape thematic map” (A. Trippe, 2015, p. 105)</td>
<td><a href="https://clarivate.com/products/derwent-innovation/">https://clarivate.com/products/derwent-innovation/</a></td>
</tr>
<tr>
<td>Questel</td>
<td>Orbit</td>
<td>“Integrated patent searching and analysis system, contains network and thematic mapping tools” (A. Trippe, 2015, p. 105)</td>
<td><a href="https://www.orbit.com">https://www.orbit.com</a></td>
</tr>
<tr>
<td>Search Technology</td>
<td>VantagePoint</td>
<td>“Text-mining and visualization tool empowered by natural language processing [...] concepts can be grouped, clustered or categorized” (Yang et al., 2008, p. 287). Visualization by matrices, maps and excel charts (Yang et al., 2008).</td>
<td><a href="https://www.thevantagepoint.com/">https://www.thevantagepoint.com/</a></td>
</tr>
<tr>
<td>Chemical Abstract services</td>
<td>STN AnaVist</td>
<td>“Online analysis tool designed to work with the STN international system, contain dynamic charts and thematic maps” (A. Trippe, 2015, p. 105)</td>
<td><a href="http://www.stn-international.com/stn_anavist.html">http://www.stn-international.com/stn_anavist.html</a></td>
</tr>
<tr>
<td>Gridlogics</td>
<td>Patent iNSIGHT Pro</td>
<td>“Patent research and analysis platform that includes specialized text mining algorithms for patent and scientific literature, powerful charting and mapping capabilities” (A. Trippe, 2015, p. 105)</td>
<td><a href="https://www.patentinightpro.com/">https://www.patentinightpro.com/</a></td>
</tr>
<tr>
<td>LexisNexis</td>
<td>TotalPatent One®</td>
<td>“Research and analyze patent data and protect intellectual property, contains user-driven text mining tools” (A. Trippe, 2015, p. 105)</td>
<td><a href="https://www.lexisnexisip.com/products/total-patent-one/">https://www.lexisnexisip.com/products/total-patent-one/</a></td>
</tr>
<tr>
<td>Linguamatics</td>
<td>Linguamatics</td>
<td>“Text mining software that can be used to analyze scientific literature, patents and other sources, features subjects, action, object triplets” (A. Trippe, 2015, p. 105)</td>
<td><a href="https://www.linguamatics.com/">https://www.linguamatics.com/</a></td>
</tr>
</tbody>
</table>

*Table 1: State of the Art in patent content analysis and mapping tools*
In 2010, Yang, Akers, Yang, Klose, & Pavlek expanded their academic contribution from solely describing many tools’ capabilities (Yang et al., 2008) towards executing case studies focused on analysis for patent mapping making use of VantagePoint and different STN tools, including AnaVist (see Table 1). Yang et al. (2010) performed two case studies with the tools; a technology-assessment related study and a company assessment related study, in order to check the performance of each tool. As “companies are often interested in profiling a technology area or an organization as part of a technical intelligence or competitive study” (Yang et al., 2010, p. 205). Yang et al. (2010) drew conclusions about the integration of the tools into patent analysis workflow, based on the users’ experience during the case studies. This research will take the assessment of the tools a step further in order to add validity. In this research is chosen to move towards a quantitative method on which a value judgement can be made based on quality measure of the tools. The quality of a software tool can be assessed by judging whether “specified goals can be achieved with effectiveness and satisfaction by specific users carrying out specified task in specified environments” (Bevan, 1995, p. 115). Similar to Yang et al. (2010) this will follow to conclusions about the tools’ integration in business processes.

Table 1 again adds proof for the premises by Grandstand at the beginning of this paragraph, due to the development in computation power and analysing power, the availability of automated patent mapping and clustering tools has expanded over the years. However, as the development in this technology domain certainly improved, the output of the automated tools are not always perceived as truthful. “Within the professional patent information community there is still a high degree of scepticism as regards the use of these new linguistic technologies. At least in part, this is due to the relative ‘black box’ effect inherently attached to the nature of the said technology” (Fattori et al., 2003, p. 335). Personal contact with IP-analysts within Philips led to a similar conclusion. To put it differently, text mining and visualization tools might be perceived as a ‘black box’ due to the lack of insight in what happens within the tool. The user of the tool can easily have doubts whether the output of the tool is indeed that what was looked for. In order to overcome the black box effect, the different tools on the market have to be assessed in their contribution to patent analysis in a real-life setting. Because, whenever a valid judgement can be outspoken about the added value of tools in companies’ processes, the black-box effect might decrease. Furthermore, this valid judgement is necessary for determining the effectiveness of a tool (Munzner, 2014) in the form of IP-based intelligence.

In automated tools it is likely that a discrepancy exists between the task the user wants to apply the tool for and the task that the tool actually executes. A reason for this is because the
users of the tool has no insight in what is happening inside the tools. Software engineers face lots of limitations during the development of tools. Software design in general, also text mining and visualization design is full of trade-offs (Baeza-Yates & Liaghat, 2017; Munzner, 2014), “and most possibilities in the design are ineffective for a particular task” (Munzner, 2014, p. 1). This discrepancy adds to both the black-box effect and the quality (Bevan, 1995) of the tool for a specific task. Therefore, next insight will be given into the technology behind automated patent clustering and mapping tools.

### 2.5 Technologies of automated patent mapping

Both the Questel (n.d.-b) and Clarivate (2018) help-pages provide an overview of the steps their tools take in the patent analysis and mapping process. Also Trippe (2015) described similar steps in his description of ‘spatial content mapping’, and he also provides a technical definition; “Mapping is related to clustering or classification exercises, where the system involved take the document clusters or classes and arrange them in 2-dimensional space by considering the similarity of the documents relative to one another over the entire collection. Documents that share elements in common are placed closer together spatially, while ones with less similarity are placed further away” (Trippe, 2015, p. 54).

“Computer based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively” (Munzner, 2014, p. 1). Clustering and classification are often used interchangeable, but are in fact two different things (Trippe, 2015). In this research is looked into mapping based on clustering. “Clustering is normally associated with unsupervised methods of organizing document collections based on a similarity comparison between documents”. This unsupervised clustering can be described as clustering without the use of any teaching signals (Tucker, 2004). In other words, in this case the unsupervised clustering in the tools uses an algorithm for categorizing patents based on their textual content without any interference of the user. Where classification is normally associated with supervised learning (Trippe, 2015), supervised learning means that examples are classified manually and the cluster algorithms learn from these manual allocations (Zeidat, Eick, & Zhao, 2004). The tools under investigation all make use of unsupervised techniques, therefore fall in the category of clustering.

Patent mapping tools perform their analysis in the following steps (Clarivate, 2018; Questel, n.d.-b; Trippe, 2015); they first build a vector model based on the extraction (and weighting) of terms in the patent text. Next, potentially clustering occurs based on the multi-
dimensional matrix. In order to map these clusters, dimensionality reduction for mapping is applied to the matrix. Dimensionality reduction can be directly applied on the vector model, then the proximity of coordinates for the mapping represent the clusters. Based on a reduced matrix, the actual mapping can occur, which results in a scatterplot with contours and colours that emphasize on clusters of points in a topographical map-like presentation. Next, these steps are explained in more detail.

2.5.1 Building the vector model

As was mentioned before, in patent mapping and clustering tools, text mining techniques are applied on (semi)-structured data fields in patent documents (Yang et al., 2008). The result of the text mining process is a vector model (also referred to as vector space model or term vector model) (Questel, n.d.-b). Term vector models were first introduced by Salton, Allan, and Singhal (1996), “in such a model, each text document is represented as a high dimensional vector” (J. A. Wise, 1999, p. 1225) in which each unique term retrieved from the text document, adds a dimension to the vector. For technology mapping, the goal is to retrieve those terms that are descriptive for technologies. There are several steps that can been taken in order to retrieve those descriptive terms for building the vector model. With combinations of the following steps text can be prepared for analysis:

- **Tokenization**: learning the computer where one word ends and another begins (Trippe, 2015).
- **Lexical normalization**: reducing words to their roots, making use of for example:
  - **Stemming**: eliminating affixes (prefixes, suffixes) to retrieve a word stem, or the root of a word (Trippe, 2015).
  - **Lemmatization**: is in some way similar to stemming, where stemming usually chops off ends of words, the goal of lemmatization is to remove inflectional endings, to return the word to the ‘dictionary form’ or ‘lemma’ (Manning, Raghavan, & Schütze, 2008).
- **Part-of-speech tagging**: The identification of nouns, verbs or adjectives (Trippe, 2015).
- **Term filtering**: “reducing the number of terms […] by removing stop words (non-content bearing terms)” (Trippe, 2015), and terms that are either to frequent or to rare (Praise Mangemba, 2018).

Also more advanced step can be taken:

- **Parsing**: “create single terms for similar words or multi-word phrases” (Clarivate, 2018).
- **Syntactic Normalization**: syntax is defined as “the grammatical arrangement of words in a sentence” (Cambridge Dictionary, n.d.-c), in syntactic normalization, sentences will be normalized to a certain word order.

- **Semantic Normalization**: semantic is defined as “connected with the meaning of words” (Oxford Online Dictionary, n.d.). In semantic normalization, words will be added that describe the meaning of a word or a set of words retrieved from the text (Praise Mangemba, 2018).

After the procedure of the above standing standardization processing on the patent documents, each retrieved term builds one dimension of the vector model. Per patent, the number of occurrences of the dimensions (terms) is counted. Optional, a weighting can also be added to a term, for example dependent on the text field from which it was retrieved.

### 2.5.2 Clustering

The clustering of patent documents can already occur on a high dimensional level. In high dimensional clustering, vectors that are most similar are grouped as a cluster. These clusters are then compared against each other and arranged in such a way that clusters showing much similarity are located close together in the high dimensional space (Trippe, 2015).

### 2.5.3 Dimensionality reduction

In text mining for 2D patent mapping, dimensionality reduction (DR) is inevitable. “DR refers to the process of mapping an n-dimensional point into a lower k-dimensional space” (Vlachos, 2017, p. 354). “The word-count vectors used to represent documents typically have thousands of dimensions” (van der Maaten & Hinton, 2008, p. 2579). “High dimensional data is notoriously difficult for humans to comprehend because of the lack of physical analogy of data with more than three dimensions” (Xu, Zhang, Pérez, Nguyen, & Ogilvie-Smith, 2017, p. 1).

The problem of analogy falls in what mathematicians refer to as ‘the curse of dimensionality’ (Donoho, 2000). The curse of dimensionality was introduced by Bellman (1957) and refers to problems that arise due to the “exponential increase in volume associated with adding extra dimensions” (Keogh & Mueen, 2017, p. 314). For example problems in analysing or organizing data high dimensional data. DR is opposed as a solution to the curse (Keogh & Mueen, 2017).

In data visualization, with DR “the relationship between the original high-dimensional objects can be visualized in two- or three-dimensional projections” (Vlachos, 2017, p. 356). Over the years, many DR techniques have been developed (Bunte, Biehl, & Hammer, 2011). DR techniques can basically be divided into two main categories; linear methods for DR and non-
linear methods for DR. Next these techniques will briefly explained and examples will be provided.

“Traditionally, dimensionality reduction was performed using linear techniques such as Principal Component Analysis (PCA)” (van der Maaten, Postma, & van den Herik, 2009, p. 1). “PCA chooses a set of representative dimensions called the principal components based on the degree of variation they capture from the original data” (Underhill, Mcdowell, Marchette, & Solka, 2007, p. 2). “In contrast to the traditional linear techniques, the nonlinear techniques have the ability to deal with complex nonlinear data” (van der Maaten et al., 2009, p. 1). Over the years, many nonlinear extensions have been developed, such as Sammon’s mapping, which emphasizes on the preservation of distances important for effective visualization (Verleysen & Lee, 2013), and Isomap, “where the distances to be preserved are based on the data distribution itself” (Verleysen & Lee, 2013, p. 619). More recent developments, such as “algorithms from the stochastic neighbour embedding (SNE) family have been shown to outperform distance-based methods in the last years, especially when the original space is high-dimensional” (Verleysen & Lee, 2013, p. 619). Stochastic techniques for dimensionality reduction preserve the probability that points in a lower dimension relate to each other. In their article, Van der Maaten et al. (2009), have investigated to what extend PCA was outperformed by twelve nonlinear techniques, and gave an indication of the major weaknesses of these twelve DR techniques compared to SNE. Verleysen and Lee (2013) described a more comprehensive overview of the state of the art in DR techniques. Besides others, these sources will help to make a value judgement on the technical design choices in the vendors’ tools.

Findings in the literature provided some possible insights into the technologies used in ThemeScape. This suspicion of the technologies used emerged out of publications by Wise et al. and Wise (1995; 1999) that where described in the literature review. Wise et al. (1995) discussed their proceedings in the Multidimensional Visualization and Advanced Browsing-project that was initiated by the Pacific Northwest National Laboratory (PNNL). In this paper, a visualisation tool called ThemeScape was disclosed to the public domain.

The following roadmap will show that it is conceivable that the ThemeScape patent mapping tool in Derwent Innovation (DI) finds its origin in the software Wise et al. created and wrote about in 1995 and 1999:

1. Working paper #94 of the CASA (Centre for Advanced Spatial Analysis) from the University College London (Dodge, 2005), describes how some PNNL’s researchers founded Cartia Inc., the company that launched ThemeScape as a product. “ThemeScape was originally developed by Cartia which in turn was a spin-off formed
by information visualization researchers at the Pacific Northwest National Laboratory (PNNL) in the mid-1900s” (Dodge, 2005, p. 10). Screenshots of ThemeScape maps included in the CASA report (page 10) show great similarities with the ThemeScape output in Derwent Innovation.

2. Bloomberg.com reports that Cartia Inc. was acquired by Aurigin Systems Inc. on February 13, 2001 (Bloomberg, n.d.-b).


4. Bloomberg.com reports that MicroPatent LLC was acquired by Thomson Reuters Corporation in 2004 (Bloomberg, n.d.-c).

5. In 2006 the ‘Intellectual Property & Science business’ of Thomson Reuters was sold and continued under Clarivate Analytics (Clarivate Analytics, 2016). As shown in Table 1, Clarivate Analytics is the current vendor of Derwent Innovation.

The current ThemeScape maps in DI are very alike with the screenshot in the CASA report back in 2005, when ThemeScape was vended by Cartia Inc. The roadmap shows that DI potentially processes its topographic patent maps on a technology described by Wise et. al. in (1999). In this publication it is described that for projection of the high dimensional document vectors onto a two-dimensional plane, multidimensional scaling (MDS) (is a class of DR techniques) is used for smaller document sets. The term MDS covers a taxonomy of algorithms that are variants of what Torgerson described in (1952), which is nowadays referred to as ‘classical’ or ‘metric’ MDS. Classical MDS is closely related to PCA (van der Maaten et al., 2009), “Metric MDS focusses on preserving distances between pairs of points” (Underhill et al., 2007, p. 2), “it finds linear transformation of data that minimizes the sum of squared errors between high-dimensional pairwise distances and their low-dimensional representatives” (van der Maaten & Hinton, 2008, p. 2595). Classical MDS and PCA assume linear relationships in the high-dimensional data. In the article by Wise et al. (1999) that describes ThemeScape, for the MDS-algorithm is referred to the work of Shepard (1962a, 1962b). Shepard’s publications, together with the work of Kruskal (1964) form the basis for what is referred to as non-metric MDS, or NMDS. This difference stems from the characteristic of the data the MDS-algorithm is applied to. NMDS makes only a few assumptions about the nature of the data (Holland, 2008), and it does, for example, not assume linear relationships in the data. Usually non-metric
dissimilarity measures are applied on data captured from text-documents (term vector models) (Ackermann, Blömer, & Sohler, 2010), so does NMDS. If linear techniques would be applied, typically, failure to preserve very similar data-points in the low dimensional (2-D) representation would occur (van der Maaten & Hinton, 2008). For the larger document sets an algorithm was developed by the authors, which was called Anchored Least Stress (J. A. Wise, 1999). Anchored Least Stress can be explained as the application of PCA (a linear technique) on multi-dimensional clusters instead of individual documents, in order to reduce computation time. For Orbit no literature was found that discloses their DR technique.

2.5.4 Mapping

After the dimensionality reduction, mapping the patent documents occurs. Each patent document occurs as a point on a map. Where the clusters are identified by labels and contours in the map. The user can add a third dimension by colouring points on the map with certain conditions, in order to generate new insights. As became clear in the above described process of automated clustering and mapping, “a huge amount of information must be cut and summarized to be useful for supporting decision-making; even though the kind of information processing could alter the quality of original data set” (Freitas et al., 2002, p. 2). Therefore, in the next chapter the methodology will be described that will be used in this research in the assessment of the automated tools under investigation, in order to make a value judgement and indicate applications as a source of IP intelligence.
3. Methodology

In this chapter, the research methodology will be described, which forms the basis for the empirical analysis. Methodology is “the plan of action, process or design lying behind the choice and use of particular methods and linking the choice and use of methods to the desired outcomes” (Crotty, 1998). In the next paragraphs, the plan of action for this research will be elaborated upon extensively. In this master’s thesis a combination of qualitative (based on data in the form of words) and quantitative (based on data in the form of numbers) methods (Sekaran & Bougie, 2013) is used in a case study design. “Qualitative research uses a naturalistic approach that seeks to understand phenomena in context-specific settings” (Hoepfl, 1997, p. 47). Which is exactly what this research entails, as there is sought for a validation of a phenomenon (automated text mining and mapping tools) in a specific context (the innovative company: Philips). However, Yin (2013) emphasizes on the relevance of quantitative data and its according methods for two reasons. In the first place, data can cover events or behaviour that is explained in the case-study. Secondly, data can relate to a certain entity that is being studied within the broader case-study. Both of these reasons hold for applying quantitative approaches to the research into content analysis of patents. Automated mapping tools are more or less an embodiment of the transition from qualitative data (text) into a quantitative visualisation (vector model and scatterplot). Therefore, the combination of qualitative and quantitative techniques is not only possible and justified, but also perceived as necessary to provide a comprehensive answer to the research question. Next, argumentation for a case study design will be presented and the case will be introduced. Furthermore, the current analysing activities at IP&S will be described. Finally, the methods that will be used for data collection and the associated research methods are elaborated upon.

3.1 Research strategy

A case study can be defined as “a research strategy that involves an empirical investigation of a particular contemporary phenomenon within its real life context using multiple methods of data collection” (Yin, 2013, p. 2). Case studies are characterized by collecting information about a specific object, event or activity. To provide an answer to the research question, one might examine a real-life situation from different angles and perspective with various methods of data collection (can both be qualitative and quantitative) (Sekaran & Bougie, 2013). In an attempt to formulate an answer to the research question, a case study as research method was
chosen, since an in-depth understanding of the phenomenon and use of automated mapping is necessary. A case-study design offers this in-depth view (Verschuren & Doorewaard, 2010, p. 178) that is necessary to answer the research question of this report. As the quality of automated patent maps depends on its performance in practice (Bevan, 1995), an empirical investigation of the mapping tools in a real life situation appears to be very useful.

The major advantage of using a case study for this particular research is that it is a practice-oriented study to gain a general picture and it is a flexible strategy (i.e. a case-study allows changes of course during the research project) (Verschuren & Doorewaard, 2010). A general picture of the object (tools) is of benefit to what Philips wants to know: whether the tools are anyhow useful for them. Furthermore, in a practice-oriented research the results are identifiable and more readily accepted by the stakeholders. Acceptance from the stakeholders (e.g. IP analysts, decision making actors) enables for the results to make a contribution to real-life change (Verschuren & Doorewaard, 2010). This change can for example be in the way the patent maps are treated or valued by the users.

3.2 Case description

As was stated before, the case-study is executed within Philips. In this paragraph the choice for having Philips as the case for this research is described. Traditionally, Philips has been a highly innovative company, this is proven by the extensive IP portfolio it holds (over 60.000 IP-rights) and the high number of patent filings each year; 1.200 new patents in 2017 (Philips, n.d.). However, also indicated by examples like the (co-) invention of well known (some of them disruptive) consumer products such as radios, televisions, electric shavers, the compact cassette, the CD, the DVD, and in other fields of technology and in non-consumer areas such as medical devices. With the generation of such an extensive base of IP, Philips as no other recognizes the importance of a well-developed IP department in a company. Not for the least part because the royalties gained from the licensing and sales of IP rights is a multi-million dollar industry. For instance, in 2017 Philips’ earnings before deduction of interest, tax and amortisation (EBITA) on IP were 225 million dollar (Koninklijke Philips N.V., 2018). Philips IP&S employs over a thousand employees worldwide (G. van der Ligt, Personal Communication, August 9). IP based intelligence is currently provided by IP&S in the form of (client-specific) IP-analysis-reports, also referred to as ‘patent landscape reports’. These reports are internally utilized by different departments in the support for strategic decision-making and IP-based policy. Executing this research at Philips opens up access to a large knowledge base on patent data analytics due to their extensive experience in IP handling in business. Currently, patent maps are not used in
their IP-analysis-reports, because for them (and many other companies) the question remains what can be learned from automated patent maps and if they indeed show valuable information. Philips IP&S is considering implementing these automated tools. An initial proposition is that such techniques are efficient and can offer new insights. As currently, multiple full-time equivalents (FTE’s) are put into the analysis and visualization of patents in order to produce IP-intelligence. Many hours are also assigned to the manual clustering of patent sets, based on the technologies the records in the set describe. Due to the huge amounts of patents existing worldwide, and the different words and languages used in patents, software tools capable of automated technology clustering by mapping could theoretically save time, money and prevent incompleteness and offer new ways of visualizing patent data. This research is about the verification of these automated maps as being a valuable contribution to the IP-analysis workflow in Philips and/or being a source of IP-intelligence.

Next, an overview of the analysis process at Philips IP&S is presented, with as an end-result a report that is shared with the requesting party within Philips. This overview is presented in order to provide insight into the current analysing activities within Philips IP&S, and to indicate where the mapping tools would fit in this process.

3.2.1 Patent analysis process at Philips

Similar to Ackoff’s (1989) wisdom pyramid, the patent analysis process starts with collecting raw data, that is transformed into communicable knowledge (in the form of a report), from which the user can generate wisdom. This process consists of five steps: Definition, Data collection, Processing, Analysis and Reporting.

Definition

The very first step towards the end report of the analysis is to define what is exactly the question to be answered. Internal clients from various Philips business entities send in requests for landscape reports towards the Philips IP&S together with the IP analyst an order definition (i.e. problem definition) will be formulated. The order definition should give answer to the next questions:

- What is the real question to be answered?
- Why is this required?
- Who wants to know this?
- What is the business impact?
- When should it be delivered?
Based on the comprehensive order definition following from the answers on these questions, the IP analyst starts the assignment by collecting the data needed for further analysis.

Data collection

Data will be collected based on the problem definition. The data is retrieved from different patent databases (Dewent Innovation, PatentSight, Orbit etc.) and can be referred to as (big) IP data. The goal of the data collection is to retrieve that exact set of patents the analyst needs in order to answer the question. Collection of the data is mainly done by writing search queries for the different databases. Writing these queries is done by IP analysts that are skilled in the art and possess in-depth knowledge of technologies that Philips is active in. In collecting data, and thus in writing queries there is usually a trade-off between recall and precision. Recall can also be described as ‘completeness’; a high recall means that a high amount of the total relevant patents in the database are retrieved. Precision can also be described as a low level of noise in the dataset. Noise is caused by data that is not relevant and can influence the results. High precision means many of total hits retrieved out of the database are indeed relevant.

Note that this research is not about the composition of datasets, but about the processing and analysis of these datasets. In the analysis made in this research the assumption is made that the IP-analysts at Philips have made the best trade-off between recall and precision and provide sets of IP data that can be relied upon and that are useful for further analysis.

Processing

The next step in the creation of the report is the processing of the dataset retrieved from the data collection. In the first place, the processing takes place based on metadata. Software tools are used to categorize or clean datasets based on, for example, assignee, families, citations, legal-status or value indicators (economical value of patents). Based on which factors the processing proceeds depends on what is looked for in the problem definition. Important to note is that processing and analysing are iterative processes that can occur alternately multiple times in one report.

Analysis

The next step is the analysis part, this is where the focus of this research lays. In the analysis step the patents are clustered based on their content. This means by looking mainly at textual fields in the patents, such as title, abstract and claims. Clustering is done in such a way that groups of patents represent technologies or applications. The clustering can be based on queries
that filter specific technologies, certain classification codes (e.g. CPC and IPC) or by human interpretation after scanning sets of patents. Especially the often unavoidable human scanning through the records is a very time consuming activity. The purpose of this research is to examine software tools that can do this clustering in an automated manner (by patent mapping) as a potential contribution to the IP-analysis process that is described here. Another motivation for this research is to gain insight in whether the maps could function directly as a source of IP-based intelligence for strategic purposes.

Reporting
The last part in the process is to convert the analysis into a report such that IP data is converted into explicit knowledge. The report usually consists of the problem definition, charts based on the processing and analysis of the data and a conclusion. Figure 5 provides a schematic representation of the IP-analysis process.

Figure 5: Visual representation of the IP-analysis process at Philips IP&S

3.2.2 Patent mapping tools at Philips
Research will be conducted into the true value of tools capable of automated patent mapping in the creation of IP-based intelligence, during a case study at the IP-analysis department of Philips IP&S. As stated in the introduction, by the courtesy of Philips, licenses were provided that enabled the use of ThemeScape (as part of the Derwent Innovation patent analysis module) and the analysis module of Orbit Intelligence for this research. Philips holds a very extensive experience in the field of IP-analysis and worked with many different tools in the past. The reason that these two tools were chosen is because, based on these experiences, Derwent Innovation and Orbit Intelligence were selected to be implemented into Philips’ IP-analysis workflows (E. Balidemaj, personal communication, July 2, 2018). Both of these tools offer an additional automated patent mapping functionality; for Philips it was therefore a logical choice.
to provide specifically these two tools for the execution of this research. Besides do both vendors of the tools hold an decades-long experience in patent analytics and have a large worldwide installed base for their database use and analysing software related to patent data (Clarivate Analytics, n.d.; Questel, n.d.-c). As key players in the market for patent data analytics, their software is supposed to be indicative for the state of the art in commercially available patent mapping tools. As show in Table 1 the direct competition is not big, around four other companies have a very comparable mapping functionality in their tool (KMX Patent Analytics, Wisdomain, VanatgePoint, STN AnaVist), the other three tools in the table are remotely equal to ThemeScape and Orbit as they have similar functionalities. Having access to ThemeScape and Orbit tools offers the possibility to research whether automated patent mapping can by any means contribute to a company’s IP-based intelligence. In the next paragraph, the rationale for the chosen methods and data sources will be elaborated upon.

3.3 Research method and data gathering

In the literature review it was found that the world of patents is not as straight forward as it might look at first instance. The field of patents is a highly complex area, in which knowledge management, law, economics, linguistics and technologies meet. The automated software assessed in this research is used to reduce this complexity, such that the content of large patent sets is easy to comprehend. The validation of this transition from a complex world to a straightforward visualisation, asks for a research method that is capable of dealing with many different aspects. This was found by the technique of triangulation. Triangulation is a technique in which the research is addressed from multiple perspectives (Sekeran & Bougie, 2013). A major strength of a case-study is the opportunity to use different sources for data collection, and opportunities for different research methods (Yin, 2013). “The idea behind triangulation is that one can be more confident in a result if the use of different methods or sources leads to the same results” (Sekeran & Bougie, 2013, p. 104). Using only one method for studying a certain phenomenon makes the study more vulnerable to errors typical for the chosen method (Patton, 2002). Triangulation of methods can be seen as a means for verification by checking consistency in findings (Yin, 2013). However, the purpose of the triangulation method is not solely verification of the results, since in this case there is dealt with a complex situation, whereby the context is very relevant. The triangulations’ major contribution to the research is that the use of different methods helps to provide deeper understanding of the phenomenon (Cohen & Crabtree, 2006). The available IP databases, analysing tools and experts at Philips offer these multiple methods and sources for triangulation. For example, the ability to process
large sets of patent data with different software tools offers ways to do quantitative research, where the presence of many experts in the field of IP-analysis offers ways to do qualitative research. Together this offers the availability to approach the problem from many different angles.

In Figure 6, a high-level overview of the research strategy with a description of the methods for triangulation is presented. This graphical representation shows how the research into the tools for automated patent mapping is fully demarcated within the triangular approach. As can be seen in the picture, the verification of these tools will occur by using multiple methods for research and data collection for the creation of an overview of capabilities, an evaluation of effectiveness and a technical assessment by introspection.

Figure 6: High-level overview of research strategy
For the purpose of making a *quality* judgement of software tools, Bevan (1995) prescribed to measure usability as a measure of quality. Usability is the “extend to which specified goals can be achieved with effectiveness and satisfaction by specified users carrying out specified tasks in specified environments” (Bevan, 1995, p. 115). The research into the usability of the tools will occur by different methods. An indication of how well the tools are capable of carrying out specified tasks in specified environments (i.e. creating IP-based intelligence for innovative companies) will be provided by an overview of each tools’ *capabilities*. An assessment of the tools’ *effectiveness* will be conducted by setting up an experiment to make this measurable. Lastly, *introspection* will generate more insight into the findings of the foregoing researches. Introspection can help the users of the tools in deciding under which conditions patent maps are indeed useful for this purpose.

In the next paragraphs, the methods and data sources for researching the tools’ *capabilities*, *effectiveness* and their software review by *introspection*, will extensively be elaborated upon.

### 3.3.1 Capabilities

The research into the tools’ capabilities is executed in order to find out how automated patent maps can be used as a source for business or competitive intelligence. In order to do this, in the first place the capabilities *of* the tools will be researched. This depends on the *functionalities* the user has within the tools. So, the capabilities *of* the tools will be assessed by their *functionalities*. Once the functionalities of the tools are sorted out, research was done into the capabilities the users has *with* the tools’ output (i.e. patent maps). The capabilities *with* the tools can be referred to as how the patent maps can be used as IP-intelligence, so their real-life *applications*. This real-life application can be referred to as what Bevan (1995) calls ‘the specified task in the specific environment’ (see quote above). Both the *functionalities* and the output of the tools, as well as their *real-life applications* are researched using qualitative methods. These qualitative methods will be described next.

*Functionalities*

Research into the tools’ functionalities was done by examining literature sources and in particular the tools’ ‘help’-documents. This was done because the help-documents are written with the purpose to help the user understand the functions in the tools and how to use it. Furthermore, online teaching sessions were set up. In these sessions, vendors’ experts on the tools showed the functions and elaborated on the processes in the tools. Lastly, personal contact
with the vendors’ support personnel helped to clarify or verify the earlier findings about the functionalities in ThemeScape and Orbit.

**Applications**

How automated patent maps could potentially function as a source of IP intelligences was in the first place researched by performing a literature review. As was observed in the literature, different scholars had already indicated the potential applications for automated patent mapping. Within Philips was sought for real-life cases to underpin these findings in the literature. Also, personal contact with Philips’ employees in different departments led to verification or indications of the applications. The description of real-life cases for the application of automated patent mapping is referred to as *case-exploration* research. With a case-exploration, examples will be provided on how automated patent maps could potentially function as a sources of IP intelligence. Next will be explained how and why the quality of the tools was assessed by measures of effectiveness and by ‘introspection’.

3.3.2 Effectiveness

The effectiveness of each tool will be assessed in an experimental setup. In this experiment, patent data that Philips knows through-and-through will be analysed by the tools and this will be used to quantify the tools’ effectiveness. The tools’ effectiveness will be measured by *accuracy* and *efficiency* of the clustering performance. The technology clustering forms the basis for all the insights that can be generated from the output; clustering performance is therefore indicative for the effectiveness of the tools.

To evaluate the effectiveness of the different mapping tools the *task* for which they are used has to be assessed. “Computer-based visualization systems provide visual representations of datasets designed to help people carry out task more effectively” (Munzner, 2014, p. 1). As was seen before, the major tasks for which mapping tools can be used are clustering of technologies and visualizing these clusters in such a way that the user can effectively identify the technology fields that are represented by the dataset. As the quality in clustering is of direct influence on all other functionalities of the maps (paragraph 4.1), the clustering performance of both tools is treated as the most descriptive quality measure in this research. Put differently, the quality in clustering influences for a large part the insights one can generate from a map. As, “a primary purpose of visualization is to generate insights” (Saraiya, North, & Duca, 2005).

In order to validate visualization-tools’ effectiveness on a specific *task*, a common approach is having a target user to try a tool and based on his findings assess whether the tool
is in fact useful (Munzner, 2014). In this research, the target user is the IP analyst and the subjects under investigation will be the IP analysts at Philips IP&S. In generalizing the results of this research, this domain of users can expand towards IP analysts in general (or even policy makers that use IP intelligence).

“Typically, visualizations are evaluated in controlled studies that measure user performance on predetermined tasks” (Saraiya, North, & Duca, 2005, p. 443). A quasi-experiment is chosen to validate the clustering-capabilities of the mapping tools. A quasi-experiment can be defined as a study that “exposes an experimental group to a treatment and measure its effects” (Sekaran & Bougie, 2013, p. 182). In order to measure the effects of using a mapping tool for the task of clustering, quantification of the results contributes to the ease of drawing conclusions. There are different researches that worked towards a quantitative method of accessing visualization. Common quantitative measures for the effectiveness of a visualization tool are efficiency and accuracy (Bevan, 1995; Chen & Yu, 2000; Saraiya et al., 2005). Timing of the experiment will provide an indication of the efficiency. Accuracy can be measured by precision and the amount of correct and incorrect responses a tool can generate in a certain predetermined task (Saraiya et al., 2005).

**Efficiency**

The efficiency in clustering will be determined by comparing the time it takes to manually cluster a set of patents with the time it takes to do it with help of automated mapping tools. Consulting IP-analysts in Philips showed that manual patent clustering occurs by clustering families based on different metadata analyses (e.g. grouping by CPC/IPC classes) and scan/read through text fields in the individual records and assign individual/groups of patents to a certain category. This process usually takes between one and three days, depending on the size of the set of patents (E. Balidemaj, personal communication, July 2, 2018). As mentioned before in the literature review, manual clustering is a time consuming activity.

**Accuracy**

Defining accuracy based on precision and the amount of correct and incorrect responses (Saraiya et al., 2005), corresponds to the way the quality of the data collection in the IP-analyses process is assessed at Philips (described in the literature review). Precision is similar to Philips internal quality measure for data collection. ‘The amount of correct responses’ is referred to as Recall by Philips internally (van der Ligt, 2017). Therefore, to quantify the results of the
experiment, the formulas for calculating \textit{Recall} and \textit{Precision} will be adapted from Philips’ processes.

Hence:

\[
\text{Recall} = \frac{\text{Relevant hits retrieved}}{\text{Total relevant in database}} \quad \text{and:} \quad \text{Precision} = \frac{\text{Relevant hits retrieved}}{\text{Total retrieved}}.
\]

However, in order to be able to calculate the recall and precision of the mapping tools, an expert evaluation of the maps is necessary. The patent clusters in the maps often do not carry the same name for a technology. Therefore, it is not always clear what the clusters in the map represents. An interpretation by an expert of the cluster names provided in the maps is therefore necessary. Besides that, the clusters in the ThemeScape maps do not offer clear boundaries. ThemeScape does provide titles of its clusters in its maps, but which patents relate to each title is subjective and has to be determined from the map. There are different ways to make a selection of those patents related to each title, and this again depends on the human interpretation of the map.

During the experiment, IP analysts from Philips were exposed to a map based on a predetermined dataset. The subject was asked to make selections in the landscapes of a given topic. The selection should consist of those patents that he or she thinks represents a given topic, based on the clusters of patents in the map that carry a title given by the tool. The analyst could make use of the different selection options in the particular tool (e.g. clicking on contours, clicking on topic-name, circle selection, rectangle selection or free form selection). The selected sets of patent numbers by the IP analyst will then be treated as the ‘cluster’ of that given topic. To give an indication of how this is executed, Figure 7 presents a screenshot of ‘free form’ selection of patents in ThemeScape.
Next, these selected clusters were checked for their accuracy by comparing them with the predetermined set of patents that comprises the topic in the initial dataset. This comparison will be done, based on the abovementioned formulas for Recall and Precision.

The ‘Relevant hits retrieved’ will be calculated by checking how many patent numbers appear in the participant’s selection of a certain topic, that also appear in the predetermined set of that particular topic.

The ‘Total relevant hits in database’ are the amount of patents that comprise the predetermined topic.

‘Total retrieved’ refers to the amount of patents in the participant’s selection. After defining these parameters, the recall and precision could be determined and together with a timekeeping of the experiment a conclusion will be drawn about the accuracy and efficiency of the mapping tools.

*Experimental design*

During the execution of this experiment, timekeeping will give an indication of the efficiency in clustering patent sets by using automated tools.
In order to make an assessment of the quality of the clusters of concepts provided by the mapping tools, recall and precision was measured as stated above. In order to do so, a dataset was needed from which it is clear on beforehand how the different patents should be grouped together if the technologies described in these patents were clustered. This is necessary because the amount of ‘Relevant hits’ should be clear beforehand in order to calculate recall and precision.

However, as became clear in the literature study, a uniform definition how patents should be grouped together is missing. The different ways of defining a patent family or patent class and the high amount patents needed to describe one technology these days (Kur & Dreier, 2013), makes it easy to justify for one patent to be part of many different groups. In this experiment it is therefore chosen to use the company’s own definition of technology clusters as a means to compare the software-based clusters under assessment. Mainly driven by the fact that the company is the actor for whom the tool was designed. One data-set that appeared to be very suitable to use for this experiment are all Philips’ active patents that are clustered in ‘Philips own classification’. Like the EPO and PCT (IPC/CPC codes) (EPO, n.d.-b), also Philips created a classification structure under which they categorize their own patent portfolio based on its business divisions. The classification consists of topics that identify certain technology fields and/or applications of technologies. The Philips’ classification is dictated by Philips business divisions. Every patent that is new to the Philips portfolio is classified according to the business division it is used in. This is done by the responsible IP-Counsel(s) for the specific division. Because this classification is structured by a group of highly skilled internal patent attorneys that are aware of Philips’ businesses, the assumption is made that this classification is the best way to structure Philips’ patent portfolio. Therefore, in order to calculate recall and precision, these classifications was used.

Philips’ almost entire patent portfolio is divided under this classification. For the experiment, all the patents in Philips’ portfolio that are divided under their own classes (around 9000 records) are uploaded into Orbit and ThemeScape (by a single patent number per family). With both tools different maps were generated based on (subsets of) this dataset. In ThemeScape this was done by analysing the DWPI title and abstract for each patent (in paragraph 4.1 this is explained and motivated), in Orbit the analysis occurred according the default settings. Instead of representing each individual patent as a single dot in the maps, for both tools only one patent per family was analysed. The reason for this is that in this way the size of the patent family is of no influence to the outcome of the map. By analysing all individual family members, bigger families would be weighted heavier in the map (i.e. would influence
the density of certain clusters). Analysing one member per family is also according to what ThemeScape recommends (see appendix A). For determining families, the tools’ own family definitions where used (DWPI-family for ThemeScape and FAMPAT for Orbit; see appendix A and B). This was done because no major differences were expected to be found due to the two definitions, and the tools work most smoothly by using its own definitions.

The Philips’ categorization is set up with main classes and sub-classes. Again, many of these sub-classes are also divided in lower classes. For the creation of the maps for the experiment, first, all the patents in Philips’ portfolio where mapped in a single map. Next, certain main classes were selected from the Philips portfolio, and also maps were created based on these patent sets representing a main classes.

The participants in this experiment are a group of three IP-analysts from IP&S in Philips. In their day-to-day work these people work on IP-analysis-reports in fields that are related to Philips’ activities. The total list of main and first level subclasses in the Philips’ classification was shown to these participants, with the question whether they could indicate those main and subclasses of which they have extensive knowledge about. Their selections were used in the experiment as the topics they were asked to identify in the maps. It is necessary that the person making the selection in the maps has knowledge about the topic he or she is selecting, since he or she can than identify all the related technologies to the topic. Whereas someone inexperienced in a certain field might fail to identify certain related technologies shown in a map, due to their lack of knowledge about it.

Both the ThemeScape and the Orbit map of the total Philips portfolio were shown to the participant. The assignment for the participant was to make selections of four given main classes, that describe certain technologies (making use of the above mentioned selection techniques), based on what was provided in each map. Also, three maps based on sets of patents that represent certain main classes where shown, with the question to identify sub-classes describing technologies. In order to check whether there are correlations between the participants’ experience in a certain technology field and the recall and precision following from the selection in the map, the participants’ own indication of their level of knowledge on the topics were added (indicated on a 1 to 7 Likert scale). Based on the results it was determined whether there is a correlation between the accuracy scores (dependent variable), and the level of the participants knowledge (independent variable) on a certain topic. With these correlations can be determined whether it makes a difference if the user is highly skilled on a certain topic or has basic knowledge about it. If no correlation is found, the outcome functions as a stronger indication that the results on accuracy (recall and precision) are independent of the skills and
perceptions of the participant, and therefore give a closer indication of the true clustering-performance of the tools.

In order to generate more insight into the ‘rightness’ of the user’s perception to what extent a certain technology is represented in a map, correlation between accuracy (dependent variable) and the participant’s perception of how well a certain technology was captured in the selection (independent variable) was assessed. Therefore, also the level of confidence the participant had about how well his selection represented the asked topic is added to the results. After each selection of a topic, the participants were asked to indicate the amount of confidence they had in how well the selected patents represented the topic in question. This was done by indicating how strongly they agreed or disagreed on the statement: ‘My selection perfectly represents the topic in the map’, indicated on a 1 to 7 Likert scale.

The potential existence of both correlations will be measured with a correlation-analysis using SPSS (Statistical Package for Service Solution). Pearson’s test is a strong parametric test for correlation (Field, 2009, p. 176). The assumption behind a parametric test is that variables follow a normal distribution and are measured according to a metric scale (Hair, Black, Babin, & Anderson, 2009). The confidence levels measured in this test are measured on a 1 to 7 Likert scale. Metric data is according to an interval or ratio scale, where strictly speaking a Likert scale is Ordinal. But, “it has become common practice to assume that Likert-type categories constitute interval-level measurement” (Jamieson, 2004, p. 1218). In an extensive research, (Norman, 2010) found that the Pearson correlation is insensitive to skewness and nonmorality, as that are basic assumptions about the results of a Likert scale. Summarizing Norman (2010, p. 630) concludes his research with; “parametric statistics can be used with Likert data, with small sample sizes […] and with non-normal distributions, with no fear of ‘coming to the wrong conclusion’”.

The dependent variables are the outcomes in terms of recall and precision. Since these are measured by percentages, these values are according a ratio scale and therefore perfectly suited for use in a parametric test (Hair et al., 2009). Next, all the dependent variables will be checked for normality making use of the Kolmogorov-Smirnov test. This test makes a comparison of the values in the sample with a normal distribute set of scores with an equal mean and standard deviation (Field, 2009, p. 144). “If the test is significant, the distribution in question is significantly different from a normal distribution (i.e. it is non-normal)”.

Lastly, it was researched whether the accuracy results may improve after a revision of an Orbit map was made, as was found that map revision is a key functionality that Orbit holds (paragraph 4.1.3). A revision of a certain map was made by one of the participants by grouping
and deletion of descriptive terms that the tool retrieved from the patent set. Measuring accuracy then occurred similarly as was described above.

This complete experiment was set up in order to assess the quality of the automated clustering done by the tools under investigation. The results of this experiment function as an indication of how well automated tools perform technology clustering. These results can support users to judge the outcome of automated map-like patent landscapes for generating insights based on these maps.

3.3.3 Introspection

In the final part of the research, the software and design choices in the tools will be investigated. The term ‘introspection’ is borrowed from psychology, which means examination of one’s soul, or can be described as self-reflection. As automated tools are often referred to as ‘artificial intelligence’, reflecting on this intelligence should give an indication on what it really entails. Knowing what is happening in the tools can help to declare certain characteristics of the output. Furthermore, it helps getting insight into the tools to make a better quality judgement about the tools and it can help the users in judging whether the output really represents what one is looking for. Introspection will be done in the first place by consulting software providers and software experts, academic literature will then be used to make an assessment of the used techniques. For verification of a certain technique a statistical test for the comparison of two scatterplots was developed.

Motivation for introspection

As was found in the literature review, automated tools can be perceived as ‘black-boxes’ in which data is uploaded, processed and presented without the users having any clue on what happened during the processing. Having insights, into what is happening in tools creating the maps, can serve different purposes. Still, the quality of the output is of major interest to the user (indicated by the measures for effectiveness above). However, having knowledge about the way the output is constructed can help the user identify when to expect higher or lower quality output. Severe outliers can occur in clustering accuracy. As the intelligence created with the patent maps can be used for different purposes, being aware of design choices and software limitations can help find better managerial applications for the maps. It is also easier for a decision-maker to attach a certain value to a map as he or she is aware of how it was created.

Furthermore, opening up these black-boxes gives an indication of what is acquired by taking licenses for the tools, as Derwent Innovation and Orbit are no free to use tools. Especially
for a highly innovative company like Philips, which holds an extensive product portfolio and knowledge base on various technology fields, the knowledge for visualizing (textual) data might also be ‘in-house’. By opening up these ‘black-boxes’, IP&S can evaluate its policy on the acquisition of tools. Current policy mainly holds that IP&S rather has a third party to deliver tools for analysing patents instead of developing such tools in-house. This policy builds on the former experienced maintenance costs (time and money) involved in designing and maintaining in-house tools (G. van der Ligt, personal communication, April 6) from previous projects. In working towards answering the research question, describing the technologies behind ThemeScape and Orbit contributes to answering this. By opening up these ‘black-boxes’, a conclusion can be drawn about how ‘smart’ the algorithms in the tools are and how well they are perceived to be able to replace or contribute to human work.

The management of a company’s CI and the creation and usage of it is becoming a bigger challenge over time. The amount of data-source for building intelligence is growing every day and so does the worldwide patent database. The majority of organizations embrace easy-to-use analysis tools whilst decision-making is the most difficult part of management these days (Rodenberg, 2007). The best analysis tools are required to be able to handle all this data and generate valid intelligence. As was earlier mentioned, patent maps are so easy to interpret that they can be perceived as a ground truth, partly because the vendors of the tools present their tools as smart and flawless. Rodenberg (2007) quoted from the British academic and politician James Bryce (1838 – 1922) to give an indication of how important it is to be informed about your intelligence;

“Three-fourths of the mistakes a man makes are made because he does not really know the things he thinks he knows”.

This quote is highly applicable for automated patent mapping; as making blunt decisions based on a visualisation can lead to mistakes. Knowing how the visualisation was created adds to the ‘really knowing’ of what is shown in the map.

Methods for introspection

With the aim for opening up these ‘black-boxes’, different sources for information retrieval were approached. The description in the ‘help-functions’ of the tools and the teaching sessions (Appendix A and B) will also be used for retrieving information about the tools’ technologies and design choices. For example, information about the term-retrieval for building the vector model can be found in these qualitative information sources. This introspection research focusses on the clustering and reduction processes in the tools. Customer support employees of
Clarivate Analytics and Questel were approached. They were specifically asked to disclose their methodology for dimension reduction, because this is assumed to be of biggest influence to quality of the tools’ output.

“A DR technique is used to map high dimensional data to two- or three-dimensional vectors, in order to display them in a scatter plot [...] it is crucial that the visualization adequately resembles the original high-dimensional data structure and distribution” (Mokbel, Lueks, Gisbrecht, & Hammer, 2013, p. 111). In other words, it is crucial that the reduced dimensions represent the multi-dimensional space in a proper way. Often DR techniques are designed with the aim of minimizing the distance distortion after a projection of the data is made (Mokbel et al., 2013; Xu et al., 2017). This implies that there will always remain some distortion after projection. As “not all the structure and relations that exist in intrinsically high-dimensional data can be faithfully represented in the lower dimensional space, and it is not clear which relations should be preserved. The application task dictates which concessions to make” (Mokbel et al., 2013, p. 111). Therefore, in the software development of a visualization tool a DR technique should be selected that is appropriate for the application of the tool (Mokbel et al., 2013). As was seen in paragraph 2.5 of the literature review, design choices had also been made for building the vector model and the potential high dimensional clustering. Other design choices that can be of major influence to the output, are the techniques chosen for the key-word extraction. However, in this research is assumed that the DR technique is of bigger influence to the output and therefore is chosen to focus introspection on DR. The literature review shows that over the years many different DR techniques have been developed. Also, many scholars have evaluated the existing techniques (Bunte et al., 2011; Underhill et al., 2007; van der Maaten & Hinton, 2008; van der Maaten et al., 2009; Verleysen & Lee, 2013). These reviews in combination with articles on individual DR techniques, offer a great source for assessing the usefulness of a DR technique in a specific application. Without going into too much mathematical details, information on the DR technique and into its earlier mentioned distortions (Xu et al., 2017) will offer the users of a patent mapping tool insights into the quality of the technique one is using. This contributes in reducing the black-box effect. Besides, information on the DR technique does contribute to the information needed for making a decent value judgment in the assessment of the tools’ contribution to the company’s processes. Lastly, insight in the tools’ DR-technique can provide the user with information that can prevent him or her from making a misjudgement of the conclusions one can base on the output of a tool.

Contact with Questel led to a disclosure of the DR-technique in Orbit. As contact with Clarivate Analytics did not provide closing conclusions about the DR technique in their tool,
the experiment, that will next be described, was initiated. In order to verify whether ThemeScape still works according the principles that were found in the literature review, an experiment was developed to verify the use of (N)MDS on small dataset (25 records) with a statistical test. As the MDS technique (Torgerson, 1952), is one of the oldest DR-techniques (Bunte et al., 2011) and therefore well-known and described, information about the general performance of MDS is widely available. This information will form the basis of the value judgement on using MDS for patent mapping.

**Experimental design**

The purpose of the experiment is verifying the DR-technique that is suspected to be used in ThemeScape. In the search to prove whether this suspected DR technique is actually used, a statistical test was set up with which a correlation can be calculated between the *mapping output of the tool* and the *mapping output of a similar analysis making use of the suspected DR*. With the calculated correlation coefficient, a statement can be formulated about; to what extent the suspected DR technique relates to the output of the tools’ mapping. If a fairly high correlation (*r > 0.5*) (Field, 2009) is found, this shows that both mapping outputs indeed relate and that it therefore is highly likely that both maps where constructed in a similar way. Correlation will be tested with the comparison of two scatterplots. “Scatterplots encode two quantitative value variables using both the vertical and horizontal spatial position channels, and the mark type is necessarily a point” (Munzner, 2014, p. 146). The first of these two scatterplots is the result of the tools’ analyses: the scatterplot shown in the topographic mapping in which each point represents a patent record from the set of documents on which the analysis was executed. This scatterplot will be treated as the *independent variable* in the statistical correlation analysis.

The second scatterplot is the result of a patent mapping analysis set up in Python. Python is an open source programming language. The Python script (a script is a collection of statements in a file) was written in order to provide a text mining, clustering and mapping tool executing similar analyses as the tool under investigation. This script holds the benefit of insight in the design choices and the possibility to easily change the design. The design choices are interchangeable via the ‘modules’ in the script. A module can be described as a file that contains a set of functions you want to include in your application (w3schools.com, n.d.). Python hosts thousands of (many third-party) modules (Python.org, n.d.). In Appendix C the code of the mapping-script is included. The output of the script is a scatterplot representing the proximity between patent documents based on their textual similarity, which is an output like the output
of the vendor’s tool. The scatterplot that is constructed with the analysing script in Python will, in the correlation analysis, be treated as the dependent variable.

Calculating the correlation between these two scatterplots will occur based on the matrixes containing the $x$ and $y$ coordinates of each point (i.e. patent) in the plot. As scatterplots are two-dimensional visualizations, the points can be represented in an $n \times 2$-matrix, in which $n$=the number of patents in the dataset. Each row represents a patent and the columns in the matrix provide an $x$ and $y$ coordinate per patent. Both matrixes that form the basis of the two above described scatterplots will be referred to as follows:

The matrix representing the output of the vendors’ tool (or independent variable) will be referred to as $P_{\text{Vendor}}$ or $P$;

$$ P, P_{\text{Vendor}} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}.$$ 

Retrieval from matrix $P$ will occur as follows; as the output of the vendor’s tool is solely an image (can be exported as a .JPEG or .PNG), the scatterplot shown in the image has to be converted into coordinates-matrix $P$. To retrieve the coordinates a ‘scraper’ was developed in Python. With this scraper, every pixel in an image will be checked whether it meets a certain condition. If so, then the pixel coordinates will be saved. For example, the vendor’s tool output can be exported in such a way that each point is coloured red. The scraper then, can be fine-tuned such that it only saves ‘red’ coordinates (based on an RGB-code). As each point on the map is represented by multiple pixels, a clean-up of nearby pixels with similar conditions, will provide a single coordinate per point on the image. The scraper script is added in Appendix D. The matrix representing the output of the tool in Python (or dependent variable) will be referred to as: $Q_{\text{Python}}$ or $Q$:

$$ Q, Q_{\text{Python}} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}.$$ 

It is very likely that in an approach to produce a clearer or easier to interpret map, the matrix in the vendor’s tool is treated with affine transformations before the map was plotted. Affine transformations can be defined as a “a class of linear 2-D geometric transformations which maps variables into new variables by linear combinations of translation, rotation, scaling and/or shearing (i.e. non-uniform scaling in some directions) operations” (Fisher, Perkins,
Walker, & Wolfart, 2003, para. 1). An assumption made in this experiment is that in the mapping tool, before the $P_{Vendor}$ is plotted as output, affine transformations are applied. Affine transformations are applied by means of transformation matrixes. “An affine transformation is equivalent to the composed effects of translation, rotation, isotropic scaling and shear” (Fisher et al., 2003, para. 2). Each of the individual transformations will be applied by multiplying $P_{Vendor}$ with a transformation matrix and for translation of the scatterplot, $P_{Vendor}$ has to be summated with a translation vector. In the case of an 2 x n matrix, the composed transformations, or affine transformation, are represented by some transformation matrix $A_{\text{transformation}}$ or $A$:

$$A_{\text{transformation}} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$$

and by some x and y coordinates translation vectors, $t_x$ and $t_y$:

$$t_x = (t_{x1}, ..., t_{xn})^T$$
$$t_y = (t_{y1}, ..., t_{yn})^T$$

making translation matrix, $T$ (Späth, 2004);

$$T = \begin{bmatrix} t_{x1} & t_{y1} \\ \vdots & \vdots \\ t_{xn} & t_{yn} \end{bmatrix}.$$  

In his article Späth (2004) describes a method for approaching the best translation vector and affine transformation matrix between two sets of coordinates. Or, as he formulates: “Let two point sets P and Q be given in $R^n$. We determine a translation and an affine transformation […] such that the image of Q approximates P as best as possible in the least squares sense” (Späth, 2004, p. 27).

Based on the possibility to approach the best fitting affine transformation between sets of coordinates (Späth, 2004), the alternative hypothesis can be formulated as follows:

$Ha$: If, in $R^2$, the DR technique used for determining $P$ is similar to $Q$, then the following should hold:

$$P \approx Q * A + T$$
knowing that both analysing methods (vendors’ tool and Python script) likely differ in design choices (besides the DR technique). For example, different designs in the text analysing part of the tool (see list of text preparation option under the heading building the vector model in literature review) or flaws in the precision of the scraper. These differences will cause noise in the comparison between the matrixes (as outputs slightly differ). Therefore, also an error: \( \varepsilon \), representing noise, is included in the equation, from which follows:

\[
P = Q \ast A + T + \varepsilon
\]

in which:

\[
Q \ast A + T = P^\circ
\]

hence:

\[
P \approx P^\circ.
\]

The null-hypothesis is therefore formulated as follows:

\( H_0: \) If, in \( R^2 \), the DR technique used for determining \( P \) is different from the DR technique used in \( Q \), then the following should hold:

\[
P \neq P^\circ.
\]

To test these hypotheses, the Pearson’s \( r \) will be calculated as a measure of correlation (Field, 2009) between the coordinates in \( P \) and the coordinates defined by \( P^\circ \). High correlation would give a strong indication that the DR-techniques used in the determination of both sets of coordinates is similar, where the unexplained relation between \( P \) and \( P^\circ \) refers to the error, or put differently:

\[
\varepsilon = r - 1.
\]

The correlation between both sets of coordinates will be checked by applying a Pearson’s \( r \) test on the correlation between the x-coordinates of the independent \( P \) and the dependent set of \( P^\circ \) coordinates. The same test will be applied for y-coordinates of \( P \) and set of \( P^\circ \) coordinates. The found \( r \)-value’s indicate to which extent \( P \) and \( P^\circ \) relate to each other.

The combination of the above described research methods and their outcomes will add to the methods triangulation approach and will form the basis for the conclusions. In the next chapter, the results that followed from the above described methodology will be presented.
4. Research and Results

In this chapter, the results will be presented that followed from the different researches in the triangulation-approach, as was shown in Figure 6. Paragraph 4.1 will show an overview of capabilities of the tools, divided in the users’ functionalities and the potential applications for the tools’ output. In paragraph 4.2 the results following from the evaluation of effectiveness will be described. In paragraph 4.3 the introspection will disclose techniques and design choices within the tools, that were assessed by a software review.

4.1 Capabilities

In this paragraph the ‘front-end’4 related capabilities of each tool are described. This means, the capabilities users have with the tools in real-life applications. To make this description, different research approaches were undertaken. Firstly, the vendors of ThemeScape and Orbit were consulted and they provided comprehensive training on the usage of their tools. Besides, regular contact via e-mail with the vendors provided a clearer view of what the tools are capable of. Lastly, working with the tools with random datasets offered also insights in their functionalities.

Mapping tools offer many functionalities to the user in order to generate new insights with the map (Clarivate, 2018; Larner, 2018; Praise Mangemba, 2018; Questel, n.d.-b). However, in order to generate trustful insights, the technology clustering in the map should be flawless, as these functionalities find their basis in how the technologies are represented in the map. For example, to provide insight in which technological fields competitors are active, the technology clusters in the map should be according to the competitors’ activities to make a valid conclusion. It has become clear in the literature review, and will become even more clear in the following chapters, that it is hard to define what is ‘correct’ in patent clustering. As the world of patents is very complex, it is hard to define clear boundaries in technology fields represented by patent clusters. Following from this a major threat in the use of these maps was stated by one of Philips’s IP-analysts:

“Automated patent maps are very approachable and easy to interpret for people, this makes that often more value is given to the insights from such maps than should have been” (F. Obers, personal communication, July 2, 2018).

4 Back and front end are terms used in fields of computing, where ‘back end’ refers to the part of the computer system where data is processed and stored and the ‘front-end’ refers to those parts of the software that are directly seen by the user (Cambridge Dictionary, n.d.-d).
Again, the way in which the analysed patents are divided and grouped together in the map is of major influence of the insights and potential (managerial) conclusions that follow from such a map. In the next paragraphs, the functionalities of each tool are described. The focus will be on the tools’ functions that influence the technology clusters provided by the tool.

4.1.1 Functionalities
In order to get a good impression of the functionalities in both mapping tools, online teaching sessions were set up. In these sessions, vendors’ experts on the tools showed the functions and elaborated on the processes in the tools. The functionalities described below are based on these teaching sessions, relevant findings described in each tools’ ‘help’-function and personal contact with the vendor’s support team. This resulted in an overview in which both tools’ functionalities are compared. For clarity in referencing and for providing a clearer understanding, a summary of the relevant parts of both training sessions are added as Appendix A and B (containing screenshots for gaining better insights into what is described).

The comparison of functionalities will occur on the bases of the options a user has in the different stages of mapping process. Those stages will first be explained. In the next two paragraphs the functions of each tool will be described in these stages and then summarized in a table-format in order to provide a brief but comprehensive overview.

The process of mapping starts after a certain set of patents is selected from the world wide databases (e.g. retrieved by search queries or by uploading a predetermined set). The first comparison is therefore made on the number of documents or records each tool can handle. The next stage is the options for analysis stage, in this stage the options are described the user has to influence the process of mapping, beyond selecting the set of documents on which the analysis occurs. The following stage is about the adjustment or revision of the processed map. In this stage is described to what extend the user is capable of making changes within the processed map in order to refine the map towards the purpose the user made it for. The last stage is about the functionalities in the map; this shows the options the user has within the map.

Further will described how the user can derive different insights from the maps. This is done by the introduction and description of some real-life applications for the tools’ output. These applications are found in literature, contact with the vendors and personal communication with (potential) stakeholders within Philips.
4.1.2 ThemeScape

Clarivate describes ThemeScape as follows: “ThemeScape is a data analysis tool that creates content maps from Derwent Innovation patent data […] a content map is a visual representation of a collection of documents organized by thematic content. This helps you analyse large data sets using a familiar metaphor” (Clarivate, 2018).

**Number of records**

Derwent Innovation offers the ThemeScape mapping functionality under the tab ‘analyse’, which can be selected after a set of patents is retrieved out of the Derwent World Patent Index (DWPI). “DWPI covers more than 34.7 million patent families from 50 worldwide patenting authorities and 2 journal sources” (Clarivate, 2018). The size of the retrieved set form the DWPI has to have a minimum of 20 records for ThemeScape to analyse and a maximum of 3 million individual records. When the family-collapse is on, ThemeScape can process a maximum amount of 60,000 records (Larner, 2018). Family collapse means that the result set only shows one member per family (Clarivate, 2018).

**Options for analysis**

The options for analysis in ThemeScape are extended. For its analysing process, ThemeScape needs at least one text field in each patent to analyse each patent of the set (Larner, 2018). There are many text fields to select, for example the title, abstract or claims in a patent. The text fields in the analysis are ‘mined’ as explained in the literature review. ThemeScape will identify terms with a technical meaning in text fields of patents. The frequency of occurrence of certain terms in a patent and the similarity of terms used between patents is the basis on which the clustering occurs.

Derwent Innovation offers value-added data as it re-writes titles and abstracts to a standard format being consistent in use of wording (Larner, 2018; Trippe, 2001). The DWPI abstract is a summary of the patent, with a focus on the claims (Larner, 2018). Therefore, Clarivate does claim in their help-document that “including DWPI data improves the quality of the resulting analysis. The consistency of language used in DWPI’s all-English abstracts provides a landscape map far superior to one using data from patenting authorities alone” (Clarivate, 2018). In further assessment of the tool in this research, this strong recommendation to use DWPI text fields in ThemeScape analysis has been acted upon. The DWPI text fields seemed to be of great contribution to the ThemeScape mapping capabilities, and therefore using these fields should best represent the quality of ThemeScape. Besides the normal abstracts,
Derwent Innovation also offers different types of DWPI-abstracts that can be used for analysis for different purposes (Larner, 2018). In Table 2 an overview is shown of the functionalities of these special abstracts in ThemeScape.

<table>
<thead>
<tr>
<th>DWPI Abstract</th>
<th>Purpose in ThemeScape</th>
<th>General Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract – DWPI Advantage</td>
<td>Focus on the benefits of the invention (Larner, 2018).</td>
<td>“Covers the advantages of the invention as described by the author” (Clarivate, 2017).</td>
</tr>
<tr>
<td>Abstract – DWPI Detailed Description</td>
<td>Focus on inventive step (Larner, 2018).</td>
<td>“Included when it is not possible to summarize the main claims of the invention within the novelty field” (Clarivate, 2017).</td>
</tr>
<tr>
<td>Abstract – DWPI Drawing Description</td>
<td>No perceived function in ThemeScape.</td>
<td>“Explains the technical drawings included in the record” (Clarivate, 2017)</td>
</tr>
<tr>
<td>Abstract – DWPI Use</td>
<td>Focus on the inventions application (Larner, 2018).</td>
<td>“Covers all the uses (applications) of the invention in terms of its different technology areas” (Clarivate, 2017).</td>
</tr>
<tr>
<td>Abstract – DWPI Tech Focus</td>
<td>Not enabled as selection in ThemeScape</td>
<td>“Describes the technology incorporated in the Invention”(Clarivate, 2017)</td>
</tr>
</tbody>
</table>

Table 2: Special abstracts in ThemeScape

There are a few more special DWPI –abstracts, those are however not included in Table 2 as those are only applicable to patents in the biological or chemical field (e.g. Abstract –DWPI Activity and Abstract – DWPI Mechanism) (Clarivate, 2017).

Per field that the users wants to include in the analysis, he or she has to select the treatment of the field in the ThemeScape analysis. There are four options: Analyse, Summarize, Both and None. Table 3 describes these functions:
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analyse</strong></td>
<td>Field used in ThemeScape’s analysis algorithms, but field is not included in the <em>popup</em> (see Appendix A for example) that displays when hovering over a patent in the map (Clarivate, 2018). Topic (term) retrieval in ThemeScape occurs by analyses making use of <em>parsing</em> (see literature review for explanation).</td>
</tr>
<tr>
<td><strong>Summarize</strong></td>
<td>Field is not included in the popup that displays when hovering over a patent in the map, but field is not used in ThemeScape’s analysis algorithms (Clarivate, 2018).</td>
</tr>
<tr>
<td><strong>Both</strong></td>
<td>“Applies both the Summarize and Analyse treatment to fields” (Clarivate, 2018)</td>
</tr>
<tr>
<td><strong>None</strong></td>
<td>Includes the field in the <em>Document Details</em> (see Appendix A for example). But, field is not included in the popup that displays when hovering over a patent in the map and field is not used in ThemeScape’s analysis algorithms (Clarivate, 2018).</td>
</tr>
</tbody>
</table>

*Table 3: Treatment functions in ThemeScape*

Trials during this research showed that including meta-data fields with the treatment ‘*Analyse*’ (e.g. Publication Number, IPC code, assignee) into the ThemeScape analyses were of no influence to the outcome of the mapping. In other words, these fields do not influence the clustering of patents when treated with ‘*analyse*’. Such field’s default treatment setting is *summarize*, for the purpose of easily and quickly reviewing these values when selecting records on the map (Clarivate, 2018). When included in further analysis, these fields will therefore be treated in their default setting.

Another option the user has to influence the analysis in ThemeScape is the list of stop words. Stop words are terms that will not be included in the analysis (Larner, 2018). By default, this is a list of words with no technical meaning, because the purpose of the analysis is to produce clusters of patents based on their technical content.

**Map revision**

After the map has been processed, it will be saved in Derwent Innovation. By selecting a saved map, the file description opens. In the file description the user can find the date, the patents that have been used to process the map (can be exported) and the number of dropped records (Appendix A) (Larner, 2018).
The peaks shown in the ThemeScape map indicate technologies with lots of activity, described by a label, the contour lines around a peak show individual documents related to that technology (Clarivate, 2018). There are two types of labels in the ThemeScape maps: peak labels and regional labels. Once the map has been processed, the terms shown in the map can manually be changed.

**Functionalities in the map**

The map shows clusters of patents per technology (replicate what a human being would do while reading them, as the peaks shown in the map represent ‘piles of patent documents’ that are categorized based on the technology). The presented technologies should provide insight in which technologies are represented by the analysed set of patents.

ThemeScape automatically generates groups on the basis of: assignee/applicant, patent status (dead or alive), countries/regions, IPC-4 codes, publication years, estimated expiration years and topics in the map (Clarivate, 2018). These groups, or parts of it, can easily be selected and highlighted (in different colours) in the map. According to Clarivate (2018) highlighting these automatically generated groups does provide instant insight in the following:

- “Which companies are working in which technology areas and where opportunities exist for new development
- Where different inventions are patented around the world and where you may have new opportunities
- Specific application or a general technology category
- How the technology landscape or a company’s focus has changed over time” (Clarivate, 2018)

Selections within the map can also be made by making use of the Derwent Innovation search screen, within the ThemeScape map then can be searched for patents in the same manner as patents can be retrieved from the DI database (i.e. data collection step of the IP-analysis-process, described in Chapter 2). For more information about this search window the reader is kindly referred to the ‘help’ document of Derwent Innovation. Selections in the map can also be made by: clicking on the contours shown in the map, rectangular and circle selection and with the implicit link function (shows related records in other regions in the map) (Larner, 2018).

Via the DEAD/ALIVE function, active, indeterminate and expired patents can be assigned with different colours in the map, in order to for example show currently more successful technology fields (Larner, 2018).
An extra functionality that ThemeScape offers is the functionality of time slices. “Time slices let you create views of the map that show how the distribution of records on the landscape changes over time. This helps you find trends in patents over time” (Clarivate, 2018). With the time slices functionality the user can select a time period, ThemeScape then shows only those records in the map that fall within the period. The date on which these time slices are based can for example be the application date or publication date.

4.1.3 Orbit
The Orbit landscape is part of the analysis module in the Orbit software. The analysis module “is a tool for decision making. It allows you to analyse big volume of data (several hundred thousands of families) in a very short time” (Questel, n.d.-b). When a certain set of patents is retrieved from the database in Orbit, the button analysis can be applied to open the analysis module.

Number of records
The Number of Records that can be analysed in Orbit is 2.000.000. Saving an analysis can only be based on 30.000 records or less (Questel, n.d.-b).

Options for analysis
The influence of the users on how the map will be processed is not extensive in Orbit. The user can, before running an analysis on a set of patents, either choose to group the set by family or to select all individual patents. Depending on this selection, each single point in the map will represent either a family (according to the family definition in Orbit) or a single patent.

Concept (or term) retrieval from the text fields in the map appear to be very advanced as Orbit makes solely use of nouns and applies lemmatisation and syntactic and semantic analysis (Questel, n.d.-b). In the literature review an overview of design choices for term retrieval in text documents (for building the vector model) is provided.

Map revision
The map revision functionalities is where the main strength of Orbit appears to lay. Via the tab: ‘data rules’ (available via all analysis in the analysis module), technology concepts extracted from the patents are shown in order of occurrences (counted by the number of documents that share the concept). Map revision can occur by deleting (and including) and grouping (and ungrouping) concepts. The revision options are also available for: Assignees, Parent companies,
Inventors, Representatives, IPC codes, ECLA codes, CPC codes and Plaintiffs/Defendants. However, as the mapping is based on text fields, excluding these field from the analysis has no effect on the map (it has effect on other analyses in the analysis module). For clearer insight into the concepts shown in the data-rules tab, the analysis module offers the ‘key technology concepts’ window. In this window all concepts are shown that were used for processing the map, behind each concept a number is shown which indicates the amount of analysed families that share that concept (Questel, n.d.-a). In the ‘clouds’-representation of concepts (see Appendix B for a screenshot) the font size is an indication of the importance of that concept within the set of analysed families (Questel, n.d.-a). “Every concept is weighted according to the field in which it has been identified, and the places where it occurs. The concepts are a results of the semantic content of the patent” (Questel, n.d.-a). There is however a difference in the concept generation in the ‘map’ and in the ‘key technology concepts’ window. It can occur that one concept shown in the key technology concepts, appears in more than one cluster in the map. This is due to the fact that the algorithm reducing dimensions might create different clusters than the clustering algorithm used for generating the key technology concepts (T. P. Mangemba, personal communication, May 28, 2018). Revising the concepts making use of the key technology concepts tab does make the algorithm for dimensionality reduction handle terms differently. For example, when deleting a non-descriptive term from the key technology concepts will make that the mapping algorithm does not take into account this term (i.e. dimension) anymore.

Functionalities in the map

A major advantage of Orbit is that it offers clear boundaries of its clusters in the map. Orbit automatically assigns labels to certain clusters in the map; these labels are the top three concept terms found in the cluster. By clicking on a label all records that are behind that label are automatically selected. Also, points falling under a specific label are coloured uniformly. New labels can easily be added by a free drawn selection of a set of records in the map. If the user wants for example to see which technologies best indicate a certain unlabelled cluster in the map, with the ‘label patents’ option the user can draw a selection and a label for the selection will be generated. A right click on a label offers a dropdown menu with other concepts that the user can select for replacing the label.

Generating insights with the map can mainly be supported by the ‘Colour by’ function. With this function the user is able to highlight points with certain values in the map. This
function reaches its limitation at nine different colours that can be assigned to one or different selections of a certain type of data (Questel, n.d.-b). These types of data are:

- Assignees
- Inventors
- Representative
- IPC, ECLA or CPC codes
- Legal status
- Legal state
- Concepts
- List or work file (a set of patents that can be uploaded manually)

By colouring records in the map, the initial colours appointed to each cluster disappear, (they return when the colours from the ‘colour by’ function are disabled). When a single record in the map is assigned with more than one colours (e.g. matches two values for which a colour is assigned), the colour last assigned will be shown in the map (Praise Mangemba, 2018).

Orbit has different means for making selections in the map; the user has the possibility to select all patents by a single button, to select patents by a free-hand selection or by an ellipsoidal shape.

Insights generated from maps are the technologies represented by a certain set of patents and the proximity between different technologies. As “every point is so positioned that it respects as best as possible the rules of proportionality and distance” (Questel, n.d.-b). In other words, it is tried to locate each record in such a way that its distance towards other points represents the amount of technological relatedness. Furthermore, the ‘colour by’- function holds the potential to give insight in how different patent meta-measures are divided over different fields of technologies. Again, the quality of the representation mainly depends on the quality of the technology clustering.

**4.1.4 ThemeScape and Orbit**

In Table 4, functionalities in both tools are summarized. In general, it appears that ThemeScape mainly offers means to influence the options for analysis, so before the mapping will occur. Orbit is more focused on an iterative process after the mapping, to let the user view the map, make adjustments in the analysis module-environment and re-run the mapping.
<table>
<thead>
<tr>
<th><strong>Number of records</strong></th>
<th><strong>ThemeScape</strong></th>
<th><strong>Orbit</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 3,000,000 individual records or 20 - 60,000 family indicators (Larner, 2018).</td>
<td></td>
<td>Maps with more than 30,000 families cannot be saved (Questel, n.d.-b).</td>
</tr>
<tr>
<td><strong>Options for analysis</strong></td>
<td>Many fields can be selected to be included in the analysis. Clustering occurs based on fields that contain text. It is strongly recommended by Clarivate to use DWPI – Title and DWPI Abstract as fields to analyse. The user can manually add words to the list of stop words before processing the map.</td>
<td>Before running the analysis on a set of patents the user can either select to group by family or to select all individual patents. Depending on this selection, the points in the map will either represent a family or a single patent.</td>
</tr>
<tr>
<td><strong>Map revise</strong></td>
<td><strong>Labels:</strong> Labels (peak or regional) can be manually changed into clearer terms.</td>
<td><strong>Labels:</strong> Labels of clusters can be changed into predetermined terms. Furthermore, selections in the map can be made to which automatically labels can be added that are descriptive for that set.</td>
</tr>
<tr>
<td><strong>Mapping:</strong></td>
<td>In an iterative process the map can potentially be improved by adding unclear or non-descriptive terms, found in the labels or topic list, to the list of stop words and then re-process the map.</td>
<td><strong>Mapping:</strong> In the Orbit analysis module, there are analyses included that can make ‘term-cleaning’ easier and accessible (‘data-rules and ‘key technology concepts’ analysis). Via ‘data-rules’ analysis the user can select terms to be excluded, included or grouped in the analysis. After this selection the map will re-process.</td>
</tr>
<tr>
<td><strong>Map functionalities</strong></td>
<td><strong>Selection and colouring:</strong> Easy exportation of selected records.</td>
<td><strong>Selection and colouring:</strong> Easy exportation of selected records. To provide insights, selections of patents can be made in the map and these selections can be coloured in any colour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clear boundaries in clusters, as clicking on a label selects everything in related to that label. Besides all</td>
</tr>
</tbody>
</table>
with no limitations in amount of selections. Besides, once a selection has been made, the list of records in that selection occurs in a window in which detailed information about each record can be shown.

Selections can be made by selecting automatically generated groups, by selection tools or by using the search function.

- **Automatically generated groups:**
  - The user can select groups in the categories: Assignee/Applicant, Patent status, Countries/regions, IPC-4 codes, Publication years, Estimated expiration years and Topics.

- **Selection tools:** Click on contours in the map, Circle/Rectangular/Free-form selection, implicit links (shows related records in other regions in the map).

- **Search function:** Opens Derwent Innovation Search window and this applies to the documents in the ThemeScape map.

Different selections can be combined and/or exported.

labelled clusters are coloured by default. Whenever the ‘colour by’ function is applied, the colours in the labels will decay.

With the ‘colour by’ function, the user is able to highlight points with certain values in the map. Nine colours are offered which one can link to a certain value, which can be selected from a dropdown menu (Questel, n.d.-b):

- Assignees
- Inventors
- Representative
- IPC, ECLA or CPC codes
- Legal status
- Legal state
- Concepts
- List or work file (a set of patents that can be uploaded manually)

When a single point in the map is highlighted with two colours (e.g. matches two values for which a colour is assigned), the colour last assigned will be shown (Praise Mangemba, 2018).
No clear boundaries in clusters, selection of a certain topic can proceed in different ways.

Whenever one record is included in two or more selections, it will colour white in the map.

**Time Slices:**

“Time slices let you create views of the map that show how the distribution of records on the landscape changes over time. This helps you find trends in patents over time” (Clarivate, 2018)

*Table 4: Overview of functionalities in ThemeScape and Orbit maps*

### 4.1.5 Applications

In this paragraph, a selection of applications for the above described tools will be described, supported with some examples from (potential) real-life cases. These applications are based on studies of literature and findings through communication with different departments within Philips. Applications for the tools as a source of IP-intelligence will be described. Also the insights that can be generated with the tools will be elaborated upon.

**Automated mapping for efficient clustering in IP-analysis**

In the creation of IP-analysis-reports, often clustering of a set of patents is required for the visualisation of technology fields described by a certain set of records. A reason for including technology clustering into a report can for example be to map out a competitive environment in a certain field of technology. This is done by clustering technologies and showing in which of these clusters competitive firms are active. Such insights can help for example in setting up strategic alliances, scouting for Merger and Acquisitions (M&A) or to identify licencing opportunities. Technology clustering can also show in which technology fields patenting activities are high, in order to detect technological trends in the competitive environment (E. Balidemaj, personal communication, August 17, 2018). Other trends might also be detected, such as patenting trend of competitive firms might disclose something about their (IP) strategies.
Since manual clustering can be a very time consuming process (the IP analyst has to read/scan through big sets of patent content) and is based on human interpretation. Using tools capable of automated clustering can be of major contribution to this process, for both the time advantage and accuracy in the clustering results. IP&S’s main concern in this automated manner of clustering is whether such automated ways of analysis are reliable. Next will be illustrated how automated patent mapping can help in the creation of IP-analysis-reports.

The IP analyst can upload the set of patents on which clustering is desired into the mapping tool. Based on the output, the analyst can check whether the clustering showed in the map makes sense compared to his or her expression about the set that was uploaded to the tool. The analyst can also scan through some patent records to judge whether he or she agrees where it is placed on the map. Next, all, or certain, sets of patents can be exported from the map by making use of the selection functionalities. Also, the labels can be adopted, or the analyst can think of new, more descriptive labels. The revision functionalities Orbit offers, are more suitable for this application. This way of clustering would most likely save a lot of time compared to the manual alternative.

Besides supporting the IP-analysis processes, the vendors of the tools propose many insights the maps can generate (see paragraph 4.1.3 and 4.1.4). The output maps could also be included into IP-analysis-reports, in which the map can function as intelligence that can be used to support generic patent or company’s strategies. Grandstand (1999) illustrates various patent strategies that can be applied based on insight generated from abstract technology landscapes. An example of such strategy is fencing, in Figure 8 shows a visualisation of the fencing strategy as it was presented in Grandstand’s book.

![Figure 8: Fencing strategy, circles represent patents. Adapted from (Granstrand, 1999).](image)
In *fencing*, series of patents (circles in the figure) are ordered in a way to block certain lines or directions of competitors R&D activities. A fencing strategy can also be used to “build a patent wall around your product’s key differentiating features” (Rivette & Kline, 2000, p. 110). Fencing activities might also be detected via patent maps, by identification of filing activity around ‘key patents’ for a certain technology (shown by clusters of points in the map). The next paragraph will further describe how patent maps can disclose valuable information about the competitive environment and the role patent maps can play in the field of M&A.

**Portfolio mapping and the competitive environment**

In the world of M&A, IP has joined the traditional motivators for becoming a key driver of corporate combinations, such as scale economics and market share growth (Rivette & Kline, 2000). Especially in technology-intensive sectors, “patents mapping can uncover M&A opportunities and strengthen due diligence efforts” (Rivette & Kline, 2000, p. 148). With patent maps, the patent portfolios of key players in a certain industry can be analysed. Different strategies can follow out of the insights generated by a certain map.

How these maps can strengthen due diligence efforts is illustrated by the following example case. Recently a patent portfolio from Kodak was up for sale, which contained over 1,000 patents. The IP-analysts of Philips only had two weeks to evaluate this portfolio before making an analysis report out of it. When it becomes known that a company is (partly) up for sale, many times this results in a request at IP&S for the patent portfolio overview in a very limited timeframe.

**Example case: Portfolio sale at Kodak: reviewing over a thousand patents in two weeks**

Recently a part of the portfolio of Kodak was up for sale, which contained over 1,000 patents. Kodak produces imaging products (with a history in photography), their patent portfolio could therefore be of interest to Philips’s imaging activities. The IP-analysts of Philips only had two weeks to evaluate this portfolio before making an analysis report out of it.

In Figure 9, a patent map is shown in which the portfolio of Philips and Kodak are represented, as they are today. In the map, Kodak’s portfolio is coloured yellow, and Philips’ portfolio is coloured blue. It is now relatively easy to see how both portfolios relate to each other, and whether acquiring Kodak’s portfolio would cope with Philips’ current business...

---

5 Due Diligence is referred to as it is used in accounting. Cambridge Dictionary defines *due diligence* as follows; “the detailed examination of a company and its financial records, done before becoming involved in a business arrangement with it” (Cambridge Dictionary, n.d.-b).
strategies. By inclusion of competitors in the different technology areas a judgement can be made on the changes of Philips’ competitive position in certain markets after an acquisition would have occurred.

**Figure 9: ThemeScape patent map: Philips and Kodak portfolio**

Based on Figure 9, actors within Philips could judge whether the Kodak patents in the right corner are of contribution to Philips’s business strategy. What also could lead from this map is further research into the small groups of yellow dots (Kodak patents) in the upper left of the field, because those patents really cluster within Philips’s technologies. Having these insights created in less than one hour indicates that such maps might contribute to ad-hoc due diligence, as this was necessary for the Kodak-case in the past.

This illustrative example is based on the current accordance between both the Philips and Kodak complete portfolios. At the time a decision had to be made to acquire a certain part of the Kodak portfolio that was up for sale, eventually Philips did not acquire it.

IP motivated M&A can for example offer patent protected entries to new markets; patent maps can help to identify M&A-candidates for this purpose. M&A can also be a means for strengthening the competitive weaknesses in a firm’s patent position, patent maps can help locate the acquisition target for these purposes (Rivette & Kline, 2000). “Another common goal of M&A strategies is to buy companies with underutilized patent portfolios that can be tapped for incremental revenues through licencing” (Rivette & Kline, 2000, p. 158). On the other hand, when a company wants to sell off business units, patent maps can help to select possible interested buyers.

In the literature review was mentioned that the management of IPRs falls in the domain of knowledge management. In the knowledge assets of a firm, also human capital is of major importance, as the inventors assigned to the patents hold extensive knowledge about the
invention that often goes beyond what is described in the patent record. Mergers and acquisitions are therefore followed by major management challenges. One thing that certainly has to be checked during due diligence efforts is whether the people working in the M&A candidate are on board as well (Rivette & Kline, 2000). As M&A is often inevitably related to human capital management challenges. In the following paragraph is described how patent maps support R&D allocation, illustrated by a case found in Philips; IP generation projects.

**R&D investment allocation and IP generation projects**

“A crucial element of formulating a firm’s technological innovation strategy is determining whether and how to protect its technological innovation” (Schilling, 2013, p. 181). For a company to be incentivized to undertake R&D (research and development) it must be able to appropriate sufficient returns to make the investment worthwhile (Levin & Nelson, 1987). Appropriability can be defined as “the degree to which a firm is able to capture rents from its innovation” (Schilling, 2013, p. 182). The appropriability of an innovation is generally determined by the ease with which competitors can copy the technology and mainly is a function of the technology itself and the means chosen to protect the innovation (Schilling, 2013). “Some technological innovations are inherently difficult to replicate. A firm’s unique prior experience or talent pool may give it a foundation of technical know-how that its competitors do not possess” (Schilling, 2013, p. 182). Especially when this know-how is tacit, which means that the knowledge is “associated with skills […] that people develop through their own experience in specific contexts and has essentially personal quality that makes it hard to formalize or communicate” (Newell et al., 2009). There are, however, often inventions and innovations that need protection in order to prevent competitors from copying. One of the means to do this is by legal protection of the intellectual property (as explained in the literature review) or in particular by patents. “A patent confers, in theory, perfect appropriability (monopoly of the invention) for a limited time in return for a public disclosure” (Levin & Nelson, 1987, p. 783). Patents, however, may not always be the right means to ensure appropriating returns. For example, in industries where product life-cycles are short, the time the process from application towards the potential grant of a patent takes (2 to 5 years), is often too long (Schilling, 2013). A certain technological innovation strategy executed by Philips that is particularly focused on appropriating via patents became clear during this case study. In this process, a role for automated technology mapping was foreseen.

There is a lot of research on going within Philips, for which strategies for optimizing appropriability have to be developed. There are also R&D projects in which the focus depends
on the outcome’s patentability and not on the innovation itself per se. These projects are internally referred to as IP-generation-projects (IPgen). The main target of these projects is generating new patents. In IPgen is specifically sought for new fields of application of existing technologies (I. Berezhnyy, personal communication, August 8, 2018). For focussing R&D activities towards this undiscovered application of a certain technology, the communication between the IP analysis department and the allocated group of researchers is of major importance. Where IP&S is supportive for diverse actors and departments within Philips, this is also the case for IPgen-projects by giving guidance in ‘what to invent’ in terms of chances for legal protection. IP analysts have to provide information about what extend ideas are eligible for patent protection, depending on what is claimed in patents or already part of the public domain (prior art). The support provided by the IP-analyst (by means of IP-analysis-reports) mainly determines which direction the researchers will pivot their activities to expand their chances of receiving a granted patent. Once granted, such a patent can offer licensing opportunities, or strengthen the company’s ‘fence’ (as in the abovementioned fencing-strategy).

An IPgen project starts with defining the scope of the project, this scope is communicated with the IP&S department in order to generate an IP-analysis-report. This report has the target of the identification of so called white spaces. White spaces are ‘gaps’ in a certain technology field that might lead to new business opportunities, as they are not described through patents yet (Siwczyk & Warschat, 2011). In other words, areas with minor patent protection that therefore permit business opportunities. The absences of patent protection on certain products or technology areas functions as the main driver for innovation in IPgen projects.

IPgen projects are focused on acquiring IP by R&D efforts and focus on finding new and innovative technology-application combinations (or more specific the lack of them: potential white space). Often these white space analyses are presented as a matrix. Typically, in such a matrix, combinations of technologies from a certain field are mapped against application fields (E. Balidemaj, personal communication, July 24, 2018). In table 5 an abstract example is showed of a white space matrix.

<table>
<thead>
<tr>
<th>Application 1</th>
<th>Technology 1</th>
<th>Technology 2</th>
<th>Technology 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application 2</td>
<td>Patent A</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Application 3</td>
<td>Patent C</td>
<td>*</td>
<td>Patent C, D</td>
</tr>
</tbody>
</table>

*Table 5: White space matrix. Adapted from Wich (2016).*
Every white space (no existing application of a certain technology) in the matrix is marked with an ‘*’ and might indicate a possible area for obtaining patent protection and therefore a potential valuable business opportunity. However, not every white space offers this opportunity by definition. There are several reasons why a white space in the matrix does not per se guarantee an opportunity for patentability and/or a business opportunity. There can for example be a logical explanation why a certain technology-application combination is not yet patented, for example because it might be not technically feasible. Besides this, patentability can be obstructed by any prior art on a found white space. Whenever it turns out there is an option for patentability of a such white space, still the effort and spending have to be made for researching into the white space. “From a company’s point of view, the main function of a patent should be to help the company recover sufficient returns from its investment when commercializing a new technology” (Granstrand, 1999, p. 184). In other words, there should be sufficient returns foreseen, to indeed invest in an IP-generation project before executing it. In the case of fencing, the foreseen protective value has to be sufficient. Determining the value of a (not yet existing) patent, or a company’s patent portfolio is challenging. Ernst and Omland (2011) developed the ‘Patent Asset Index’ which is a benchmarking methodology that offers an accurate assessment of a firm’s patent portfolio compared to that of the competition.

A function foreseen for automated patent mapping in support for scoping R&D activities in IPgen-projects is in the form of clear communication. Automated patent maps can support in the communication between the R&D personnel and the IP-analyst, for the retrieval of the right set of patents for the generation of the white space matrix (table 5). Often these IPgen projects are on a highly specialized field of technology, about which the researcher has very extensive knowledge. The IP-analyst however, has to retrieve all related patents from the worldwide patent databases for the scoping of the project. Via automated patent maps, the IP-analyst can share his findings in a visual manner with the researcher. The researcher can indicate related fields in the map that are worth further investigation, on the basis of which the IP-analyst can retrieve more specific patent sets. This can result in an iterative process between the researcher(s) and the IP-analyst. After which a highly specific technology – application matrix can be built in order to identify a white space. Currently, the scoping is done by the allocated team of researchers, by reading through all the patents in an IP-analysis-report on a certain topic. Making use of patent maps might improve the quality and speed up the process of scoping.
The above three paragraphs show examples of how patents can play a role in competitive strategies and R&D activities. Again should be emphasised on the fact that the support for such strategic managerial decisions should be of perfect quality since such decision can have major impact on the company’s performance. Therefore, patent maps must be trustful. In the next paragraphs, the output of the tools will be assessed. In paragraph 4.2 the tools’ output will be assessed by its clustering performance (effectiveness of the tools). Paragraph 4.3 broadly elaborates upon how the ‘back-end’ of the tools is constructed and this analysis functions as bridge between the user and the software developers, which can give a better indication to the user about applications for the output.

4.2 Effectiveness
In order to make an assessment of the effectiveness of ThemeScape and Orbit, an experiment (see paragraph 3.3.2) was set up. More specifically, the target of the experiment was to assess ThemeScape’s and Orbit’s clustering quality through a measure of effectiveness. As mentioned before, the effectiveness can be assessed by measuring efficiency and accuracy (Bevan, 1995; Chen & Yu, 2000; Saraiya et al., 2005). In the next paragraph the results of these two measures will be described and emphasized upon.

4.2.1 Results

*Efficiency*
During the execution of the experiment, a measure of time was used to assess the efficiency in clustering by the tools compared to manual clustering. The time it takes to cluster technologies in an automated manner is the combination of the time to generate the map and an identification of clusters in the map (see paragraph 3.3.2). During the experiment these times were collected per map, and are presented in Table 6. Each set of patents on which a map was generated has a different amount of predetermined classifications in it. In manual clustering there is generally sought for clusters to divide the entire set in, and no predetermined number of clusters is available. The dataset used in this experiment does have this predetermined clustering as this was necessary for executing the experiment (for calculating recall and precision). The participant was asked to identify a fixed number of technologies in the map. Therefore, the way of automated assisted clustering in the experiment is not completely equal to how manual clustering usually appears in patent analysis processes. However, the timing results shown in Table 6 give a good indication of the fastness of automated clustering by the use of technology
maps compared to manual clustering. This was according to the expectations, because in manual clustering the reading through patents takes most of the time, where especially that part is automated in the tools. The conclusion can be drawn that automated clustering is in fact very efficient. In order to judge whether is indeed as effective as the vendors of the tools claim, question remains how accurate the clustering is. During the experiment, accuracy was also measured and in the next paragraph these results are provided.

<table>
<thead>
<tr>
<th>Map</th>
<th>Number of Records</th>
<th>Themescape</th>
<th>Orbit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Analysing</td>
<td>Selections*</td>
<td>Total</td>
</tr>
<tr>
<td>Map 1</td>
<td>+/- 10,000</td>
<td>5 min.</td>
<td>60 min.</td>
</tr>
<tr>
<td>Map 2</td>
<td>+/- 2000</td>
<td>5 min.</td>
<td>45 min.</td>
</tr>
<tr>
<td>Map 3</td>
<td>+/- 300</td>
<td>2 min.</td>
<td>20 min.</td>
</tr>
<tr>
<td>Map 4</td>
<td>+/- 500</td>
<td>5 min.</td>
<td>10 min.</td>
</tr>
</tbody>
</table>

Table 6: Time measures

* The time indicated by selection is the average time the three participants took on making selections of equal topics in the same map.
Accuracy

The outcomes of accuracy in terms of recall and precision of the experiment will be presented next. The results of the calculations of recall and precision based on the selections in the different maps made by the participants are shown in table 7. In the table, the main and sub classifiers will be referred to in the results as ‘main 1, 2, 3…etc. and sub 1, 2, 3…’ as these classifications are considered confidential. Per topic in each map the average of all participants (total=3) that made that selection is included.

<table>
<thead>
<tr>
<th>Map</th>
<th>Topic</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ThemeScape</td>
<td>Orbit</td>
</tr>
<tr>
<td>All patents</td>
<td>Main 1</td>
<td>78%</td>
<td>61%</td>
</tr>
<tr>
<td></td>
<td>Main 2</td>
<td>68%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Main 3</td>
<td>53%</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>Main 4</td>
<td>81%</td>
<td>52%</td>
</tr>
<tr>
<td></td>
<td>Main 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sub 1</td>
<td>82%</td>
<td>72%</td>
</tr>
<tr>
<td></td>
<td>Sub 2</td>
<td>39%</td>
<td>49%</td>
</tr>
<tr>
<td></td>
<td>Sub 3</td>
<td>62%</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>Sub 4</td>
<td>45%</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>Main 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sub 5</td>
<td>33%</td>
<td>58%</td>
</tr>
<tr>
<td></td>
<td>Main 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sub 6</td>
<td>86%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Sub 7</td>
<td>89%</td>
<td>82%</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean:</td>
<td>65%</td>
<td>67%</td>
</tr>
<tr>
<td>Median</td>
<td>Median:</td>
<td>68%</td>
<td>66%</td>
</tr>
</tbody>
</table>

Table 7: Measures of Recall and Precision

In Table 7 all the results in terms of recall and precision are summarized by calculating the mean, which is the average of the all the scores above. As the mean has the disadvantage of being influenced by extreme scores (Field, 2009), the median of the results is shown as well. The median of a dataset is “the middle score when scores are ranked in order of magnitude” (Field, 2009, p. 21) and “is relatively unaffected by extreme scores on either end of the distribution”. When there are many similar values in a dataset accompanied with some outliers, the median might give a better representative value for the dataset.

For interpreting the above standing values in user-setting, both the mean and median are descriptive values that have to be taken into account while looking at the map. A technology
clustering in the automated map will in general have an average recall and precision according to the means presented in Table 7. Where it is slightly more likely to receive a map in which technologies are clustered with a recall and precision according to the median values shown in the same table, because the median value is not affected by occasional outliers. However, the user has to be very aware that there can occasionally occur severely low outliers in recall and precision results. On the other hand, these outliers do also occur on the positive side. Philips holds an internal baseline of 80% for both recall and precision, based on this number, the conclusion can be drawn that the clustering performance in both tools is insufficient. Each result is coloured in such a way that the darker green the box is, the better the result (in terms of the 80% baseline). On the opposite, the more red a box is, the lower the result.

In order to illustrate the recall and precision scores in Table 7, Figure 10 below is a snapshot of the selection by one of the participants of main class 11 in the map that was made out of all Philips’ patents in ThemeScape. The points coloured red is the selection by the participant, the points coloured blue is the predetermined clustering and the white points fall under both conditions. The ratio between the white (relevant hits retrieved) and blue points (total relevant in database) is a visualisation of the measure for recall. The ratio between the white (relevant hits retrieved) and red points (total retrieved) is a visualisation for the measure of precision.

![Figure 10: visualisation of recall and precision after participant’s selection of a certain topic in ThemeScape](image)

**Correlations with accuracy results**

Next, the results will be described of the test’s for correlation between the participant’s confidence in selection (independent variable) and accuracy (dependent variable) and the correlations between accuracy (dependent variable) and the knowledge level of the participant.
about a topic (independent variable). The dependent variables in this test for correlation are the outcomes in terms of recall and precision, since these are percentages, these values are according a *ratio* scale and therefore perfectly suited for use in a parametric test (Hair et al., 2009). Making use of the Kolmogorov-Smirnov test, these results will be checked for normality. Because a normal distribution of the test-results is a necessary condition for a applying a parametric correlation test. The results of the Kolmogorov-Smirnov test for each dependent variable are presented in screenshot of SPSS-output below.

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Kolmogorov-Smirnov Statistic</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThemeScape_Recall</td>
<td>0.153</td>
<td>21</td>
<td>0.200*</td>
</tr>
<tr>
<td>ThemeScape_Precision</td>
<td>0.222</td>
<td>21</td>
<td>0.008</td>
</tr>
<tr>
<td>Orbit_Recall</td>
<td>0.111</td>
<td>21</td>
<td>0.200*</td>
</tr>
<tr>
<td>Orbit_Precision</td>
<td>0.096</td>
<td>21</td>
<td>0.200*</td>
</tr>
</tbody>
</table>

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

As described in the methodology, a significant result in the Kolmogorov-Smirnov does indicate non-normality, which means that three out of the four independent variables do not meet the second condition for parametric test; normal distribution of data, therefore Pearson’s test for correlation is no longer directly applicable.

“Correlation coefficient: *r*, is a non-parametric statistic and Spearman’s so can be used when the data have violated parametric assumptions […] Spearman’s test works by first ranking the data, and then applying Pearson’s equation to those ranks” (Field, 2009, p. 181). Next, the Spearman’s Rho analysis will be applied for the different indicators of confidence and measures of accuracy (recall and precision) for both the results in ThemeScape and Orbit.

The results of the correlation test between the participant’s *confidence in selection* and both *recall and precision* are shown in Table 8 and 9. In the results from ThemeScape there are no significant correlations found. The correlations with *recall and precision* are also small. *Knowledge level* about a topic shows correlation levels of near zero and are insignificant. Both independent variables will in this research not be treated as explanatory for accuracy, as they are not significant.
### Table 8: Spearman’s correlations in ThemeScape

<table>
<thead>
<tr>
<th></th>
<th>Confidence in Selection</th>
<th>Knowledge level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td><strong>.334</strong></td>
<td><strong>.104</strong></td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td><strong>.129</strong></td>
<td><strong>.645</strong></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td><strong>.172</strong></td>
<td><strong>-.093</strong></td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td><strong>.444</strong></td>
<td><strong>.680</strong></td>
</tr>
</tbody>
</table>

For Orbit, the same conclusion applies to the Spearman’s correlations between confidence in selection and the measures of accuracy as both correlations are low and insignificant. As is also the case for the correlation between Knowledge level and the measures of accuracy.

### Table 9: Spearman’s correlations in Orbit

<table>
<thead>
<tr>
<th></th>
<th>Confidence in Selection</th>
<th>Knowledge level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td><strong>.064</strong></td>
<td><strong>.051</strong></td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td><strong>.784</strong></td>
<td><strong>.825</strong></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td><strong>.207</strong></td>
<td><strong>-.115</strong></td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td><strong>.369</strong></td>
<td><strong>.618</strong></td>
</tr>
</tbody>
</table>

Both analyses above show that no significant correlation can be found between the independent variables and the dependent variable. The correlations in both ThemeScape and Orbit between the confidence in selection and recall and precision show slightly higher $r$ values, that might explain some correlation. However due to the lack of significance the conclusion follows that there is no correlation found, and therefore the users confidence in how a selection made in the map represents a certain topic, does not say anything about the accuracy of that selection. The lack of correlation between the knowledge level of the user and the level of accuracy adds to the robustness of these accuracy findings. As the level of knowledge (beyond basic knowledge of a topic) apparently does not explain any percentage of recall and precision in the user’s selection, this recall and precision has to described solely to the performance of the automated clustering in the maps.

In the next paragraph, Orbit will be used to illustrate an example of a clustering process as it might occur in IP-analysis. The target of this next research is to give an illustration of the
options for revision that Orbit offers to improve quality of the map. Two topics that created ambiguity during the experiment were selected for this illustration and an attempt was made to clarify these topics by their appearance in the map.

4.2.2 Improving accuracy for clustering

Something that became clear during the above stated experiment is that ambiguity can exist about the difference of a technology and an application in the use of the tools. This will next be illustrated by an example: During the topic selection in one of the maps in the experiment the technology-application ambiguity came to play. In a map on a certain set of patents, the participants were asked to identify both X-Ray and CT (Computer Tomography) technologies. As CT apparatus make use of (besides others) X-ray technology (Goldman, 2007), it was hard to determine where to exactly separate these topics. One might argue that CT is an application of X-ray technologies. However, one might also argue that CT is a technology on itself. Again we come across a matter of definition, with no absolute truth. Due to this ambiguity, this map appeared to be good case for a map-revision, to see whether it is possible to separate two related topics in a map. With an aim to improve accuracy in the selection of topics in the map.

Making use of the data rules in the analysis-module of Orbit for exclusion, inclusion and combination of terms, an IP-analyst with extensive experience in the area of CT technologies (both as a researcher and an IP analyst) was asked to make a revision of the Orbit map containing CT and X-ray patents. The participant was asked to excluded those concepts from the analysis that where non-descriptive for any technology and those concepts that where descriptive for both X-ray and CT technologies, in the assumption that this clean-up would improve the separation of X-ray and CT related patent groups (and their labels).

In an evaluation about the process the participant mentioned that he had the impression that the revision of the maps did contribute to the separation of certain clusters into more descriptive clusters. So, for clustering purposes this might be a good way to speed up the IP-analysis process. The influence the user has on the clustering, seems to give the user the ability to separate certain clusters in such a way that it narrows their scope (E. Balidemaj, personal communication, July 17, 2018).

Based on the user’s findings the conclusion can be drawn that the map revision added to clearness of the labels for distinguishing technologies. As Orbit generates clear boundaries in its clusters (as all related patents are selected in the map by clicking on a label and the uniformity colours per category), this improvement in clearness of the labels holds the potential to increase the clustering performance of the map. Therefore, the insight generation by map
functionality will be increased and the maps can be of significant use for clustering tasks in IP-analysis processes. Partly due to the remained efficiency, as revision of the map took the expert around 20 minutes per iteration. But, the major concern in the revision of the maps is if it really improves accuracy of clustering, therefore for every iteration the recall and precision of every selected topic are calculated. These results are shown in the graphs below:

Figure 12: Revision of Orbit maps

Figure 12 shows that recall decreases after almost every revision, Precision seems to increase when comparing the initial map to the second revision. Both accuracy measures show fluctuation results between every revision. So, even though the clusters became clearer to the user (see above), the same improvement cannot consistently be found in the measures of accuracy.

4.3 Introspection

In this paragraph the results of the endeavours on the introspection of both ThemeScape and Orbit will be presented. As mentioned in the methodology description, this introspection research will focus on disclosure of the DR-techniques in the tools. The description in the ‘help-functions’ of the tools and the teaching sessions (Appendix A and B) already opened up information about the term-retrieval for building the vector model. In paragraph 4.1 was described that Orbit makes use of advanced techniques such as semantic and syntactic standardisation and ThemeScape uses parsing on the standardized DWPI-records. In the next paragraphs the results a provided of the research into both tools’ DR-techniques.
4.3.1 ThemeScape

Consulting Clarivate Analytics did not provide any indication about the DR-technique used in the ThemeScape tool. Since the ThemeScape technology changed owner multiple times (see roadmap in literature) and the development of automated technique developed over the years, it is not unlikely that the ThemeScape software has also changed over the years. It can therefore be that ThemeScape is not operating according to how it was described by Wise (1999) anymore. However, analysing datasets of different sizes with ThemeScape, still shows a change in visualisation by an increase of analysed records, as is represented by the differences in scatterplot appearance in Figure 13. This change in plotting appearance indicates that ThemeScape still makes use of two different DR-techniques, one for small and one for large patent sets (respectively NDMS and Anchored Least Stress).

![Figure 13: ThemeScape scatterplots as output of the ‘scraper’. Number of records processed: A = 551 and B = 1222.](image)

A and B show clear differences in spatial layouts density (Munzner, 2014). The difference seems to occur as a result of an increase of the analysed records. In map B, the output seems to have a space-filling character in the representation of a higher amount of analysed records.

“Space filling layout has the property that it fills all available space in the view” (Munzner, 2014, p. 174).

In order to verify whether ThemeScape indeed still works according the principles that found in the literature, the experiment as described in the methodology (paragraph 3.3.3) was executed. With this experiment a statistical test would show high correlation between different scatterplots (generated by the ‘mapper’ and the ‘scraper’), if they where made with a similar DR-technique.

---

6 Most likely, the Clarivate Analytics did not want to disclose their DR technique because they perceive this as the unique selling point of their tools and treat it therefore as a trade-secret.
Results of the statistical test

After an equal set of 25 patent documents was uploaded in both ThemeScape and the ‘mapper’, coordinate sets $P$ and after calculations $P^*$ were retrieved. Correlation was checked between both sets of coordinates. The result of Pearson’s test indicated no strong correlations between the independent x coordinates and the dependent set of x and y coordinates. Neither a strong relation was observed between the independent y coordinates and the dependent set of x and y coordinates. A final check for correlation between the two scatterplots is done by checking the correlation by the all the Euclidean distances in both maps. The Euclidean distance is “the distance between any two points in space corresponds to the length of a straight line drawn between them” (Pbarrett.net, 2005, p. 2). In each map, for every point the distance to all other points on the map will be calculated. Both sets of Euclidian distances are then checked for correlation. In doing this, every distances between two points have to be compared to a distance measure in the second points representing the same patent records. The resulting r value from this test is a non-significant; $r = 0.02686$. As there was no correlation found, it could not be verified that ThemeScape still uses NMDS (in the analysis of smaller record sets) by the statistical test. This does however not lead to the conclusion that the algorithms must have changed. On the contrary, there remains a high suspicion that ThemeScape still works according to what Wise et al. (1999) described in their publication. This suspicion finds its origin in roadmap described in the literature review, but also insights that were generated from working mapping script, which will be elaborated in paragraph 4.3.3.

4.3.2 Orbit

Contact with Questel’s customer support has led to a disclosure of the DR and related plotting techniques in Orbit. A variant of Force Directed Graph Drawing (FDGD) is applied to Multidimensional Scaling (MDS), after high dimensional graph drawing had occurred (T. P. Mangemba, personal communication, July 7, 2018). FDGD is not a means of dimensionality reduction, but is applied for the matter of aesthetics (Hu, 2006), it however does have a major effect on where each record is placed in the map. In Orbit the application of FDGD most likely leads to the more clearer presentation and grouping of the clusters in the map. Typically, the FDGD algorithm leads to more clarity in clusters, as related points are placed relatively close towards each other and unrelated points are placed relatively far apart.

A force-directed algorithm models a graph through a physical systems with bodies (points) and 'forces' between them, the final placement of the bodies in the map is a result of
minimization of the energy between the bodies (Hu, 2006). An explanation that speaks more to
the imagination is that in FDGD, bodies typically have repulsive forces towards each other, but
neighbouring bodies, or the edges between them, (points related to each other by for example
their textual relationship) attract to each other. The forces working on the points are visualized
in Figure 14, the red arrows represent the repulsive forces between individual points and the
green arrows represent the attractive forces between related points (indicated by the line
between them). In an iterative process, the algorithm searches for a solution in which the
‘energy’ in the map is minimized.

![Figure 14: Visual representation of the ‘forces’ in Force Directed Graph Drawing](image)

**4.3.3 Evaluating DR-techniques**

What became clear during working with the coordination matrixes of different datasets in
Python, was that the outcome of the NMDS algorithm can be very *inconsistent*. Even when
mapping the exact same data set different times, the outcome may change. This inconsistency
is also noticed in the output of the tools. When mapping equal datasets multiple times over time
(several days apart) the outcome of the map can appear differently. This holds for both
ThemeScape and Orbit. This inconsistency can be one explanation for the lack of correlation
found in the statistical test on the ThemeScape output (other possible explanations for the lack
of correlation will be emphasised on in the discussion chapter).

Inconsistency appears to be very indicative for the use of NMDS. For Orbit it is clear
that a variant of MDS is used as technique for DR because Questel shared that information. As
again data retrieved from text is typically treated as non-metric (Ackermann et al., 2010), most
likely Orbit uses the NMDS variant of MDS as a technique for DR.

For illustration, the inconsistency of the maps is shown in Figure 15 by two ThemeScape
(a) and Orbit (b) outputs, pictures A and B are the exact same dataset mapped 5 days apart.
Figure 15-a: Dataset in the field of medical imaging mapped for illustration of inconsistency: A and B show maps of identical datasets created using ThemeScape, 5 days apart.
Figure 15-b: Dataset in the field of medical imaging mapped for illustration of inconsistency: A and B show maps of identical datasets created using Orbit, 5 days apart.
The location of the clusters towards each other is clearly different in images A and B in Figures 15-a and b. The contours of the map differ as there are different amounts of hill and sea areas in maps A and B in both figures. Also, the distribution of topics in the maps is very different and even different labels appear. How this inconsistency finds its origin in the techniques for DR is illustrated in the subsequent paragraph.

**Inconsistency in Orbit and ThemeScape**

The inconsistent character of the mapping output is typical for NMDS because the algorithm contains an iterative search for the ‘best possible solution’, which can also be referred to as the *optimization problem*. “Most other ordination methods [clustering] […] result in a single unique solution to a set of data. In contrast, MDS [NMDS] […] iteratively seeks a solution and stops computation when an acceptable solution has been found” (Holland, 2008, p. 1). Therefore, this may not result in an unique solution and subsequent analysis of equal datasets are likely to result in a different clustering (Holland, 2008). In order to illustrate this iterative process, some steps are now briefly explained, without going into mathematical details.

The NMDS algorithm strives to preserve the multi-dimensional (non-metric) *distances* (user defined distance measure) between the data points in a two dimensional representation (i.e. ordination space) in with the distances between the points are measured by their *Euclidian distances*. The iterative process starts with a more or less random guess about how the multidimensional data points should be ordinated on a two dimensional surface. Next, it makes a calculation of how much all mutual distances between the data points in the two dimensional ordination deviates from the mutual distances between the data points in the multi-dimensional space. This deviation is referred to as *stress*. Stress can thus be seen as the amount of mismatch between the distances in the multi-dimensional data points and the distances in the two dimensional representation of it. The amount of stress is indicative for the ‘goodness of fit’. There are multiple ways to calculate stress, a common method was described by Kruskal (Kruskal, 1964). Next, it starts an iterative process, in which it tries to minimize the stress, by making minor changes to ordination of points in the two dimensional space. When it finds no improvement after new iterations, such that the stress level does not decrease, that point will be treated as an *optimal* representation of the multi-dimensional data set in a two dimensional space. Major drawback in this, and a conceivable cause for the inconsistency, is that this *optimal point* could be a *local minimum* of stress while typical a *global minimum* of stress is desired. NMDS “can fail to find the true best solution because it can become stuck on local minima, solutions that are not the best solution but are better than all the nearby solutions” (Holland,
In Figure 16, a representation of the global and a local minimum are given. The global minimum is that point where the measure of stress is the lowest for all the possible iterations. However, up front it is by definition unknown how much iterations are necessary to find this global minimum. Whenever the algorithm does find the lowest point of stress in its immediate neighbourhood, it is not certain whether this is a local or the global minimum. Predetermined parameters can stop the algorithm at a certain minimum, however, the global minimum might not have been reached.

![Figure 16: Visual representation of local and global minimum](image)

When analysing the same dataset again at another moment in time, the algorithm might start with a different initial guess, and another minimum can be found and treated as an optimum. It can thus happen that the algorithm stops its iterations (or optimizations) when it finds a local optimum. In a new analysis on the same dataset, the algorithm iterative process might lead to another local optimum, or to a global optimum.

In FDGD, the above illustrated optimization problem also occurs; here the algorithm tries to minimize the ‘total energy in the map’ by iterations of ordinations of points in the map. FDGD can therefore also suffer from an inconsistent character. The combination of NMDS and FDGD is believed to be the cause for inconsistency in Orbit. The use of NMDS is believed to be the cause for inconsistency in the small dataset maps in ThemeScape. According to Wise et al. (1999) PCA is applied on larger datasets in ThemeScape. PCA typically finds the one global optimum (minimum measure of stress) of dimensionality reduction, the observed inconsistency
of these maps remains undeclared. The use of PCA in processing large datasets in ThemeScape would however declare the ‘space filling’ character found in these maps.

\textit{Space filling in ThemeScape}

The space filling character of ThemeScape maps in which large sets of records where processed (shown in Figure 13-B) is the result of a major drawback in the algorithm used for dimensionality reduction. Typically linear methods, like PCA do not perform well at modelling high-dimensional data that is not linearly distributed in space (van der Maaten & Hinton, 2008)

“and it focuses on preserving the distances between widely separated data points rather than preserving the distances between nearby data points” (van der Maaten & Hinton, 2008, p. 2594). While in the mapping of clusters, it is usually better to preserve a low dimensional representation of very similar data points close together, instead of representing their dissimilarities. This dominance of the preservation of dissimilarities between points is most likely the reason why in large ThemeScape maps, such as presented in Figure 13, image B:

- No very clear separation of clusters can be determined when contours are removed from the map (making use of the scraper).
- The points fill up the entire space, as many points are represented in such a way that they should not be close together. This potentially also leads to clustering of points on the outer axes of the figure.

The authors already acknowledge these shortcomings in their description of the design choices. As they mention this typical drawback in the application of linear techniques to textual data;

“In MDS, fitting all of the pairwise distances among documents means that small deviations among points are placed at high relative importance. Under ALS, it is the large deviations that are considered important. Small scale differences in the projected placement of points onto the second plane [two dimensional plot] are sacrificed somewhat in order to arrive at a better and faster overall large scale solution. Since do much information is necessarily lost in compressing high-dimensional spaces down to the 2-D plane anyway, it seems as if this is a worthwhile price to pay to overcome a major computational bottleneck to the visualisation process” (J. A. Wise, 1999, p. 1228).

Contradiction to these negative statements about the application of linear DR techniques on textual data come from Underhill et al. (2007). They described the proceedings of research in which they checked the performance of five DR techniques on text mining. Their results show that what they refer to as ‘simple techniques’ such as PCA and MDS, can be highly effective in reducing textual data without losing important properties (Underhill et al., 2007).
However, looked at recall and precision results of the former experiment (as they not reach the desired 80% baseline), a better DR technique for patent mapping seems to remain desirable.

**Overview**

In order to provide a clear overview, Table 10 contains the advantages and disadvantages of DR and the related techniques (FDGD), that are believed to be used in ThemeScape and Orbit tools. This follows from the above described introspection.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThemeScape</td>
<td>Non-metric Multi-dimensional Scaling (NMDS)</td>
<td>Source of inconsistency</td>
</tr>
<tr>
<td></td>
<td>Preservation of distances of non-metric data points in lower dimensions</td>
<td>Optimization problem</td>
</tr>
<tr>
<td></td>
<td>Anchored Least Stress</td>
<td>Space filling</td>
</tr>
<tr>
<td></td>
<td>Fast</td>
<td>Weak preservation of close related data points</td>
</tr>
<tr>
<td>Orbit</td>
<td>Non-metric Multi-dimensional Scaling (NMDS)</td>
<td>Source of inconsistency</td>
</tr>
<tr>
<td></td>
<td>Preservation of distances of non-metric data points in lower dimensions</td>
<td>Optimization problem</td>
</tr>
<tr>
<td></td>
<td>Force Directed Graph Drawing (FDGD)</td>
<td>Clearer separation in clustering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High computation time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimization problem</td>
</tr>
</tbody>
</table>

*Table 10: Overview of advantages and disadvantages of DR (and related)-techniques in ThemeScape and Orbit*
5. Discussion

5.1 Evaluation
A case-study research was found to be the best approach for answering the twofold question of this research. The research question was on the one hand about the quality of automated patent mapping and on the other hand about the usability of patent maps, as being sources of IP-based intelligence. This two folded question about a complex technology (automated patent mapping) in an at least as complex field of expertise; IP analysis, asked for a combination of different research approaches. Therefore, triangulation of methods and data collection was used in this research design. Also, the combination of both qualitative and quantitative techniques added to the completeness of findings in the phenomena under investigation. In paragraph 5.1 will be discussed where the academic contribution of this research lays. In chapter three of this report the rationale was given for the selected methodologies in this research, paragraph 5.2 will reflect on each of the sub-researches that formed the method-triangle (as shown in Figure 6).

5.2 Academic contribution
The academic contribution of this research lays in the field of patent based intelligence and more specifically, in the automated transition of (big) patent data into actionable intelligence for businesses. By the assessment of two tools vended by major players in the market, an indication is given on the current state of art in automated patent mapping. Based on the assessment, an indication about the quality and usefulness of such tools for businesses is provided. The literature study showed that different scholars made an assessment of commercially and non-commercially available software tools for IP-analysis. However, a rigorous validation of the performance was lacking in all of them. As an extension to the current academic knowledge base, in this research a rigorous validation, by quantification of effectiveness, was made. This technique can also be used for the assessment of many other tools (even in the development phase). This research builds on the need for verification of the tools’ quality as the basis for any research into the contribution of a certain tool in business processes.

In the experiment that was setup for quantification of effectiveness, the human interpretation of the representation of certain clusters was quantified in an accuracy measure. This way of measuring automated clustering performance could be of interest for future researches by other scholars in the assessment or validation of visualization tools. Also, tool developers might use this assessment in the development and marketing of their tools. On the other hand, it can be a
quick assessment for the software licencing firm to select between certain tools (the dataset used in the assessment can be selected to the most suitable one for the perceived application of the tool).

Further academic contribution of this thesis report is found in the research methodologies that are used. The triangulation of methods approach has led to an assessment of the tools by combining qualitative and quantitative methodologies. Such as the execution of an experiment and an introspection on software tools. This introspection comprises the application of a statistical test on graphics. The application of a statistical test for the comparison of visualisations was not found well described in the literature, so the research design for determining the DR-technique can be of interest to other scholars. The research design describes how correlation can be checked between two scatterplots.

Moreover, this research can be seen as an incentive to increase the quality of automated patent mapping in general. The findings from the introspection indicated the use of techniques that have been around for decades, such as MDS and PCA, next to more advanced techniques such as FDGD and semantic techniques for term extraction. No claim is made that newer techniques will outcompete these older ones, but recent techniques such as t-SNE (van der Maaten & Hinton, 2008) show high potential.

5.3 Limitations and suggestions for further research
This thesis report conducts a case study at Philips, to generate insight into the quality and application of automated mapping tools. As a result of the method triangulation, a comprehensive answer could be formulated to the research question. However, the limitations of this research should be acknowledged and suggestions for further research will be considered.

A drawback of a case-study design is that due to the fewness of cases it is difficult to generalize findings to a broader population. Therefore, a case-study design holds the potential to lack in external validity (to what extent the results can be generalized to a broader population) (Verschuren & Doorewaard, 2010). Case studies are thus typically less generalizable than quantitative approaches. Yin (2013) however, described a dichotomy of generalizability: analytical and statistical generalizability. Case studies are analytically generalizable, not statistically. As the goal of a case study is not to generalize a small (set of) case(s) to a larger population (statistical generalization), but to generalize theoretical propositions to establish a logic that might be applicable in other cases (Yin, 2013). Limitations for generalizability of this
case-study lays in the assessment of only two tools, Table 1 shows that there are more tools out there that can be assessed in order to make a better indication of the current state of the art in automated patent mapping. However, the tools that are used are vended by major players in the market, no certainty can be given that Orbit and ThemeScape are indicative for all tools. Furthermore, the case in Philips is not by definition descriptive for other companies, so executing research into automated patent mapping applications in different companies operating in different fields should contribute to a more comprehensive picture.

This thesis focussed on IP-based intelligence, with solely patent data as a source for competitive or business intelligence. There must be acknowledged that there are many other sources on which a company can built its intelligence upon, that should be used in order to provide a complete picture. Besides, in competitive measures based on patents, the value of certain patents can also be of major importance. In this research not much attention was given to the valuation of individual patents or portfolios. Follow-up research could be done in how value measures could be included or combined with automated patent mapping. Next, the different research approaches that followed from the method triangulation will be discussed.

5.3.1 Capabilities
In the assessment of the tools’ capabilities a distinction was made between the functionalities and the applications of the tools. The functionalities of the tools were described using genuine sources found by the vendors of the tools (support employees and help documents). The applications of the tools was researched by a case-exploration. A literature study showed potential applications for patent maps and within Philips was sought for real-life case to underpin these findings. Another method to research the real-life applications of patent maps could be by in-depth interviews with decision-making actors both within and outside Philips. Via this way, a clear picture could be generated about how the actual users of IP-based competitive and business intelligence perceive the usefulness of the patent maps. However, this thesis research is mainly focussed on the validity of patent maps more than on the best applications of it. Therefore, a literature review on the potential applications of the tools that are underpinned by real-life case descriptions and expert’s opinions found in Philips was found to be sufficient for clarifying the applications of patent maps. Also, the literature review indicates how patent maps can function as a valuable source of IP-intelligence, whenever the quality is found to be ‘good’.
5.3.2 Effectiveness

The measures of recall and precision are very descriptive for the validation of automated clustering performance, partly because recall and precision fall into the IP-analyst-jargon. So, communicating accuracy in these terms is like ‘speaking in the same terms’ as the users. However, the results of the recall and precision measurers are highly depending on the predetermined clustering of the used dataset. For the experiment in Philips, a thoroughly known (by Philips) dataset was used, so the results are very descriptive for Philips. However, by assessing automated tools with more datasets that hold a predetermined classification (for example, portfolio classifications of other companies), a stronger claim could be generated about accuracy performance. Besides this, the results of the experiment in this research are based on 33 selections by in total three participants; expanding the number of maps, clusters and the number of participants would also contribute to the validity of the experiment. Lastly, defining a technology remains hard, so there will remain leads for discussion on the definition of a technology and the patents related to that definition.

The results of the experiment can be used as a support for generating insights from the map. These recall and precision numbers should be taken into account while looking at or working with the map. The results should not be interpreted as a guarantee for clustering performance. As this quasi-experimental setup is a broadly used approach (Bevan, 1995; Munzner, 2014; Saraiya et al., 2005) for assessing visualization tools, it is not the strongest setup in terms of validity (Sekaran & Bougie, 2013). Also, because there are subjective factors that inevitable affect the outcome of the experiment. These factors are the dataset and predetermined categorization of the set that was used. However, the dataset used in this experiment was carefully chosen, there will never be an absolute truth in how a set of patents should be clustered to represent a certain technology field (as became clear in the literature review). This dataset was chosen because this set finds its bases in Philips’ business divisions. Besides that, it is maintained and updated by a team of patent attorneys that are experienced in the technology fields related to the set of patents. Which means that a group of experienced people in the field of IP reached consensus on how a certain set of patents should be divided into classifications. Therefore, this classification of technologies is assumed to be superior to those clusters that are made manually by one person. Major drawback of the used classification is that it is a classification that represents certain company’s activities. Another drawback is that, due to the complexity of the field of patents (see literature review) no classification ever guarantees a best way of representing the technologies described in a patent set.
5.3.3 Introspection

The researches executed for the introspection have led to some disclosures of design choices in the tools’ software. Also, conclusions were drawn on design choices that were found highly likely to be used, however the uses could not be proved. An attempt to verify the use of MDS in ThemeScape with a statistical test did not result in a strong correlation, and therefore it could not be verified. However, this lack of correlation might find its origin in the typical character of the NMDS algorithm, as this algorithm suffers from the optimization problem. If the different attempts of checking the correlation between the mapper coordinates and the scraper coordinates did not occur on similar minima, it is likely that no correlation could be found. There are however other possibilities that potentially caused the lack of correlation found in the comparison of the scatterplots. The first reason can be that over time the ThemeScape software indeed has changed and another means for DR is used nowadays. Another reason is that in the statistical test the effect of the error (\( \varepsilon \)) was underestimated. As was described, the error represents the different design choices made in the mapper-script compared to the script in ThemeScape. An example is the term-retrieval, as this occurs very differently, the vector model on which the DR was applied might be also very different, likely resulting in different 2D plots. Additional research could be done into creating better understanding of the different design choices in vendor tools, for optimization of the Python script.

The cause for the inconsistency in the maps was allocated to the ‘optimization problem’ of the techniques. However, no declaration for the inconsistency in the Anchored Least Stress algorithm could be indicated yet. This leaves an opportunity for further research into the causes of inconsistency of this particular linear technique (or into the class of linear techniques) on textual data. The optimization problem was assumed to be the cause for inconsistency; what must be listed is that for both FDGD and NMDS there have been solutions suggested for reducing the optimization problem (Mair, Borg, & Rusch, 2016; Shepard, 1962a), however these suggestions are often accompanied with extensive computation time (Fasham, 1977; Fruchterman & Reingold, 1991). The optimization problem was found to be the most likely cause for inconsistency; this was not verified so follow-up research could be done into finding the cause of inconsistency.
6. Conclusion

In this final chapter, conclusions will be drawn based on the results as presented to the reader in Chapter 4. First, in paragraph 6.1, conclusions are given based on each research from the triangulation approach, which altogether will lead to the formulation of a comprehensive answer to the research question. In paragraph 6.2, the managerial implications of this thesis research will be described. In the final paragraph, recommendations towards different actors related to automated patent mapping are given.

6.1 Conclusion

The endeavours during the execution of the thesis research were all in service of one common goal: formulating a comprehensive answer to the research question: *What is the quality of automated patent mapping and how can it be used as a source of competitive or business intelligence?* The answer to this question will be built on the conclusions that can be drawn from the above described findings that followed from the multiple researches that were executed in the ‘triangulation of methods’ approach. The quality of automated patent mapping was assessed by a research into the usability (Bevan, 1995) of ThemeScape and Orbit, which are tools capable of automated patent mapping. These two tools were found representative for the current state of the art in the field of commercially available software tools.

In the assessment of the functionalities for real life applications was found that ThemeScape offers most functionalities for influencing the map on beforehand of the mapping, such as selecting which text fields should be included into the analysis. Orbit offers more and user friendly possibilities to revise the map that was initially processed, such as the grouping and deletion of concepts in the map. Where in Orbit the key-word extraction appears to be more advanced (semantic and syntactic normalisation) than that of ThemeScape (parsing), the standardisation in the re-written title and abstracts in the DWPI database that ThemeScape can use for analysis is likely to add to its clustering performance. Since patent text is often very complex, this standardisation in descriptive wording should help in the clustering of patents that describe related concepts. In theory, real-life applications are foreseen in the clustering process during IP-analysis and for the maps to be a source of IP-based intelligence. Automated patent maps could be applicable for M&A scouting, fast due diligence and in the field of R&D scoping.
The effectiveness (as a measure of quality) of the tools was tested by applying the tools on a dataset that Philips knows through and through. By means of an experiment, the effectiveness of clustering was measured by accuracy and efficiency. Accuracy is indicated by a measure of recall and precision. Recall measures averaged for ThemeScape at 65% and for Orbit at 67%. Precision results in Orbit were averagely measured at 66% and ThemeScape performed slightly better in precision with an average in test results of 73%. All accuracy measures suffered from occasional outliers. Besides that, the results are below the internal 80% baseline of Philips. However, the tools are highly efficient as they reduce clustering time from days to hours, or even minutes. Orbit (up to 50 minutes) has in general al longer runtime than ThemeScape (up to 5 minutes).

The results of the introspection showed that both tools suffer from optimization problems and therefore the user does not know whether the optimal representation of the data is shown. Orbit seems to have an advantage in visualisation as their Force Directed Graph Drawing creates clear clusters and the separations between them. The application of linear techniques for dimensionality reduction on large datasets by ThemeScape causes major disadvantages for display, such as space-filling and unclear clusters. Technically (term retrieval and mapping), Orbit outcompetes ThemeScape, although ThemeScape uses standardized datasets which probably adds to the overall performance.

This altogether leads to the following end-conclusion which comprises the answer to the research question: Due to the technical limitations, inconsistency in mapping and fluctuating performance of accuracy, automated patent mapping tools can only be seen as effective and highly efficient in the support of the creation of competitive / business intelligence rather than being a source of it itself. The mapping output is too susceptible to inconsistency for being a genuine source of IP-intelligence, the same holds for the clustering performance. The general quality of automated patent mapping can thus be perceived as inconsistent.

Philips’ main question entailed to what extent automated tools are an addition to the clustering tasks in their IP-analysis workflow. On the bases of the results can be determined that neither tool (ThemeScape and Orbit) can function as a replacement for manual clustering. The reason lays in the inconsistency of both mapping appearance and accuracy results (severe outliers). The space-filling characteristic of ThemeScape is outright a bad representation of clusters. Due to the space-filling characteristic only the contours shown in the map give clustering information, not the groups of points in the map. Therefore Orbit was found to be most useful as an addition to Philips’ analysis workflow. Orbit makes use of more advanced techniques and its map revision is very user friendly. A role for Orbit is foreseen in the ‘high-
level categorization’ (initial clustering of a set of patents, before scanning/reading through them) of patent sets. In general, the ‘artificial intelligence’ the tools comprise, appeared to be not very ‘bright’ and the current state of art in automated patent mapping is lacking quality to live up to their intended functionalities.

6.2 Managerial implications
The intelligence used by the higher management and decision-making actors in an inventive company can be supported by the usage of automated patented mapping. However, the quality of the patent maps does not reach the baselines (80% recall and precision). Therefore, mainly explorative insights can be created that can give indications where to increase explorative activities. Such as initial scouting for M&A opportunities, after which deeper research can be done into potentially interesting opportunities. Besides the explorative insights, automated patent mapping is very fast, which comes in very handy in the often highly competitive and rapidly changing environments of technology (Downes & Nunes, 2013). Especially in a need for fast IP related due diligence, patent maps indicate focus for further investigation very fast. A major drawback is found in the map’s consistency. However, the ‘changing’ maps are somewhat alike. Having different representations of the exact same dataset makes decision-making based on these maps vulnerable for mistakes, because the decision-making actor can never know which representation is closest to reality. This reduces the added value of these patent maps to a company’s intelligence tremendously. Best application of automated patent tools is foreseen as a support for clustering in IP-analysis, because the IP-analyst is usually skilled in the subject and can therefore make a judgement to what extend the maps can help in clustering. This human check appears to be necessary. Besides is seen that there are also some clusters that remain more or less the same in different iterations of mapping, but only appear in different places of the map. For purely clustering activities, the location of clusters towards the other patents in the set is unimportant. Furthermore, management challenges will inevitable originate after patent maps have led to an actual merger or acquisition. As the explicit knowledge (spelled out or codified) captured in patents is very often inseparably linked to the tacit knowledge of the inventor(s), challenges in human capital management will occur.

6.3 Recommendations
Based on the researches executed during this thesis project, a major recommendation is that in the use of automated patent mapping (particularly ThemeScape and Orbit) caution must be
preserved, as there is still a lot of room for improvement in clustering accuracy and the uses of the technologies. Due to the efficiency of the tools, applications in fast due diligence and clustering should definitely be considered. Orbit seems to be technically more advanced (in both term retrieval and graph plotting) compared to ThemeScape. The latest however, has the advantage of the opportunity to use rewritten and standardized text-fields to analyse (i.e. the DWPI).

Major recommendation towards the vendors is, to provide the user with a measure of stress, as this is indicative for the goodness of fit of the high dimensional data into the two dimensional representation. By accompanying the output map with a goodness of fit value, it informs the user on how well the initial dataset is represented, so this is also indicative for how reliable the maps are. Kruskal (Kruskal, 1964) already proposed such a goodness of fit number, together with rules of thumb for the interpretation of this number.

Furthermore, there could be looked into more state-of-the-art dimensionality reduction techniques; a promising example would be t-SNE. As it is particularly strong in preserving and visualising clusters (keeping related data points close, or preserving local structure) and at the same time visualising how each cluster relates to one another with clear and unambiguous ways (preserving the global structure) (van der Maaten & Hinton, 2008). t-SNE also gives its global optimum as output. Moreover, research into optimization of the keyword extraction, particular for patent documents, is key for the definition of technologies in maps, as this remains such a complex field. It might be worthwhile for a highly innovative company like Philips (already active in high-tech visualisation fields) to make an attempt in developing a patent mapping tool, in order to overcome the disadvantages of the tools that have been assessed.
References


Praise Mangemba, T. (2018). *Analysis Training by Tanya-20180502 0739-1 [webex recording]*. Retrieved from [https://emea01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fquestelorbit.webex.com%2Fquestelorbit%2FlDr.php%3FRCID%3D9f018cde20e811c3e1c4bab8d0601861&data=02%7C01%7Cjaap.vande.vorst%40philips.com%7C1ed3e4c528434289ea6308d5e00fff2b%7C1a407a2d76754d1](https://emea01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fquestelorbit.webex.com%2Fquestelorbit%2FlDr.php%3FRCID%3D9f018cde20e811c3e1c4bab8d0601861&data=02%7C01%7Cjaap.vande.vorst%40philips.com%7C1ed3e4c528434289ea6308d5e00fff2b%7C1a407a2d76754d1)


Appendixes

- Appendix A: Training ThemeScape
- Appendix B: Training Orbit
- Appendix C: Python script for patent mapping
- Appendix D: Python script for image scraper
Appendix A: Training ThemeScape

Title: Derwent Innovation training-20180418 1208-1 [webex recording]
Date: 18 April 2018
Duration: 56:26 minutes
By: Brian Larner, works for 30 years in Clarivate and for 11 years as support of all Clarivate’s patent automation products, expert on Derwent Innovation.

This document summarizes what is discussed in the Webex recording that is accessible via this link:
https://clarivatesupport.webex.com/clarivatesupport/lsr.php?RCID=06f16cf430c1d214d65e295e10d35f39

Summary
With ThemeScape you can analyse up to 3 million individual records, and with family collapse on: 60,000 records (which means one member per family). In creating the ThemeScape map, the text of the patent will be analysed. It is recommended to use one patent per family in the analysis. Because when analysing the standardized DWPI title and abstract several members of the same family can have the same title and abstract which gives higher weighting to bigger families compared to smaller families.

After selecting a set of patents for analysis, there are the following functions in the window ‘Create ThemeScape Map’:
- **Properties**: Here you have to entre a unique title, you can write a description (e.g. summary of the information that the map represents).

![Figure 17: Screenshot of the 'Create ThemeScape Map' window, tab: Properties](image-url)
- **Fields Options**: This is where you can choose which fields you want to include in your map.

![Screenshot of the 'Create ThemeScape Map' window, tab: Field Options](image)

In mapping we are looking for the words that describing the invention overall. The DWPI abstract is a summary of the patent, especially focused on the claims. DWPI is more consistent in their use of terminology (as different companies may word their technologies different even if they might be similar) therefore it is recommended to select DWPI text fields for analysis.

In analysing text, ThemeScape uses just one language and it does not translate terms. So, ensure using patents in just one language (as it otherwise forms different clusters around similar words in different languages).

Different types of DWPI abstracts can be used to analysis for different purposes (e.g.: field: Abstract – DWPI Use will focus more on the application of the invention, Abstract – DWPI Detailed description + Abstract – DWPI Novelty focuses more on the inventive step).

Per selected field to analyse you can select it ‘treatment’:

- **Analyse**: not used to create the map
- **Summarize**: not used to create, but shows the text when hovering over the patent
- **None**: not used to create the map

The order of the fields you selected has no influence on how the map is created. There is only one way where it has influence: making time slices in the map (create cross sections in the map). The field that is used for time slices is always the one that is on top of the chosen fields list (it indicates this with: *(time slice)* behind the term).

**Map Setup Options**: This tab has a list of ‘stop words’, which are terms that will not be included in the analysis. Terms that do not have a technical meaning. After you have processed the map, and you see clusters around words that are of no help in defining
technologies, you can add this word into the list of stop words and when you re-analyse the set, also this word will be ignored (process can be iterative).

- **Topic frequency, lower threshold**: Used to determine how often a topic has to occur to be analysed. Selecting 0% means that every word that occurs would be used for the analysis, it would basically include everything. Selecting 20% means that only words that occur in 20 to 80% of the patents will be used to analyse.

![Figure 19: Screenshot of the 'Create ThemeScape Map'-window, tab: Map Setup Options](image)

After you have click save, the tool will analyse the set of patents and will create the map.

Under ‘Saved Work’ in Dewent Innovation you can select the map you created. In the description of the map the number of items (analysed patents) and the number of dropped records is shown (next to owner and dates). Dropped records normally lack data that was chosen to analyse (those fields selected in the tab ‘field options’). If this is not the reason why the record was dropped, than probably after removing all stop words there was nothing left to analyse.
The processed map looks like a landscape with hills in a sea. The hills shown in the map represent clusters of patents that deal with the same type of technology. The two words occurring on the hills are the two most occurring terms in the patents found on that peak. As you move away from the centre of the peak, the patents are still generally related to the peak but do also include elements of other peaks on the map. When you select the ‘I’ button (see Figure 21 in the top right corner) and then click on a word shown in the map you can see additional terms listed on that peak (show the top terms that are occurring in the patents on that peak). You can also change the definition of peaks if you have better definitions.
Figure 21: Screenshot: example of a ThemeScape map
The colour of the peaks is also significant: white coloured peaks are the highest ones, grey are medium height, green are the very low peaks and the sea area are the patents that did not really cluster at all.

Ways of selecting groups of patents in the map (the button to select is highlighted in the screenshots shown below):

- You can click on a contour and you can export the subset for further analysis.

  ![Screenshot: contour selection](image)

  **Figure 22: Screenshot: contour selection**

- Circle function: Identify the precise set of documents that are most similar to the one in the middle of the circle (most analysed terms in common).

  ![Screenshot: Circle selection](image)

  **Figure 23: Screenshot: Circle selection**

- Implicit links function: The arrows around the circle identify the ten patents that are most similar to the patent in the centre of the circle, other than the ones that are in the circle. Arrows pointing at other areas in the map that have the strongest overlap (as patents are multidisciplinary there can be overlap with other technologies). This can show for example alternative uses for the patent or the same uses but with a different technology, therefore people working in the areas identified by the arrows may well have interest in the technologies described in the patent in the middle of the circle (represents licensing opportunity).

  ![Screenshot: Implicit Links](image)

  **Figure 24: Screenshot: Implicit Links**

- The window: ‘Document Summary’ shows the patents in the selection and the topics in these patents.
Figure 25: Screenshot: document summary (on the bases of a contour selection)
Use cases for the map:
- Automated review process for the results search by clustering patents into distinct sub-technologies (replicate what a human being would do while reading them, as the peaks represent ‘piles of patent documents’ that are clustered).
- DEAD/ALIVE functions: alive patents potentially show more successful fields
Figure 26: Screenshot: DEAD/ALIVE function
- Time slices: Comparing older patents with newer patents can be done by colouring patents published in a certain year.

- You make selections in the map making uses of the Derwent Innovation search screen, (e.g. CPC classes, assignee, keywords) by clinking on the magnifying glass.

- You can make selections in the maps based on predefined groups in the menu left in the ThemeScape window (e.g. assignees).
Appendix B: Training Orbit

Title: Analysis Training by Tanya-20180502 0739-1
Date: 2 May 2018
Duration: 135 minutes
By: Tanya Praise Mangemba, Support & Client Training.

This document summarizes what is discussed in the Webex recording that has relevance to the clustering in Orbit Landscapes. The video is accessible via this link:
https://questelorbit.webex.com/cmp3300/webcomponents/widget/playback.do?siteurl=questelorbit&recordID=124774922&isnbr=true&ticket=4832534b00000004352974d6b9c1122b3bf40bb9f6ab44e6ba2b2c2b84c8347c45ac19feafc6676a&timestamp=1530703028090&serviceRecordID=124774927&nbrhomepageurl=

Summary
In mapping, concepts are extracted from a set of patents and they are weighted. This occurs as follows. In the text, first noun phrases are identified and taken out; this means all text that is not relevant to the search itself. Next, lemmatization is applied. With lemmatization the extraction of terms is based on the main roots of the word (standardize conjugations of words). The next step is syntactic normalization, for example change “surface of screens” to “screen surface”. The final step in extracting concepts is semantic normalization, in this process also words will be added that describe the meaning (or that relate to) of a word extracted from the text. After extracting there is also scoring of words. Concepts in the independent claim are weighted more important than concepts that are found in the general claim. Weight depends on where the concepts are found in the patent and the number of occurrences. After the concepts are extracted from a set, those concepts are selected that are descriptors for the set. This is done by deleting all concepts that are neither too frequent neither too rare. Next the concepts will be clustered and groups will be labelled. Following, the reduction of dimensions will occur and next the drawing of the map itself.

Within the map, the proximity between points indicates how far off those two concepts are or how similar they actually are. Concepts that you find in the map that are close together are more related than those that are far apart. You can only compare distances between concepts in the same map, not between different maps.
Figure 28: Screenshot of an Orbit map
In the charts menu: in ‘Key technology concepts’ is shown what the main concepts that were used in processing the map are. In the tab ‘key technology concepts → ‘data rules’ you can easily exclude concepts found by the tool that don’t say anything to you from the analysis and the tool will re-group. It is good to remove certain concepts in the first 50 or 100 concepts. You can also group concepts that are describing the same thing.

For concept extraction the algorithm takes in consideration the: title, abstract and independent claims. In patent searches this can be found in Orbit under ‘Key content’; this key content is the same for the entire family. Concepts extraction is also based on the entire family (so not one member per family). Concept extraction is always based on the translated version (English text).

Once the map has been processed you for example colour by the different assignees (giving every assignee a single colour). This will colour each and every single family on who the assignee actually is. This can help to see on which technology a certain assignee has more patents than another assignee.
Figure 31: Colour by function in Orbit
You can also colour your own set of patents in an Orbit map by uploading a list of patents. Colouring certain selections can only be done up to a maximum of 9 different sets.

**Labels**
The labels that are shown are the top three concepts. You can select other concepts as well as label (via a dropdown on the label). When you click on a label you get to review all the patents that are behind that group.

You can also add labels to selections of patents in the map, by hitting the I button you can make a selection and the concepts that are behind that selection will show up in the map.

*Select all button*: When you select all the data, you see that your selection consists of less patents than the set you started with. This is because some concepts have been excluded in the key-word clean up, if there where patents which content only consisted of the concepts that had been excluded this patent would not appear on the map anymore.

When you close the analysis and open it again you always get the original map. If you want to keep the additions to the map you have done with the button: ‘save current map view’.

You can process the map based on either ‘FAMPAT’ in which one point on the map represents an entire family or it can be based on ‘FULLPAT’ in which one point in the map represents a single patent from the family.

You cannot change the name of the concepts directly, but you can delete the textbox and add a new one.

*Colouring*: When a single point is match two different highlights (falls into two categories for which a different colour is assigned), the show the last colour.
Appendix C: Python script for patent mapping

# I. Necessary Imports
import numpy as np
import pandas as pd
import re, nltk, spacy, gensim

# Sklearn
from sklearn.decomposition import LatentDirichletAllocation, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import GridSearchCV
nlp = spacy.load('en', disable=['parser', 'ner'])
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')

# II. File Handling
file = "path"
fh = open(file)

# III. Useful Functions and Objects
def sentences_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append(" ".join([token.lemma_ if token.lemma_ not in [\'-PRON-\'] else \'' for token in doc if token.pos_ in allowed_postags]))
    return texts_out

def sentences_to_text(sentences):
    new_string = ''
    for item in sentences:
        new_string=new_string+' '+item
    return new_string

vectorizer = CountVectorizer(analyzer='word',
    min_df=1,  # minimum reqd occurences of a word
    stop_words='english',  # remove stop words
    lowercase=True,  # convert all words to lowercase
    token_pattern='[a-zA-Z0-9]{3,}',  # num chars > 3,
    max_features=10000,  # max number of uniq words )

# IV. Text Processing
patent_dict = {}
all_docs = []
all_ids = []
with open(file) as fp:
    for n, line in enumerate(fp):
        line = line.strip()
        if (n==0):
            header = line
            next
        else:
            if (n==1):
                header = line
else:
    try:
        alist = []
        content = line.split("\t")
        my_id = content[0]
        title = content[1]
        abstract = content[2]
        sentences = abstract.split(" | ")
        sentences.append(title)

        # simple cleaning of words and punctuation
        processed = list(sentences_to_words(sentences))

        # lemmatization; standard word forms
        lemma = lemmatization(processed, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])
        all_lemma = sentences_to_text(lemma)
        all_lemma = all_lemma.lstrip()
        all_docs.append(all_lemma)
        all_ids.append(my_id)
    except IndexError:
        pass

vectorized = vectorizer.fit_transform(all_docs)

# V. Matrix Handling
dense_format = vectorized.todense()
row_sum = dense_format.sum(axis=1)
norm_matrix = dense_format/row_sum

# VI. Modelling
import sklearn.manifold as skl
import matplotlib.pyplot as plt

mds = skl.MDS(n_components=2, n_init=1, max_iter=100)
X_mds = mds.fit_transform(norm_matrix)

x = []
y = []
for point in X_mds:
    x.append(point[0])
    y.append(point[1])
plt.scatter(x, y)

# add AN to every Point
for n, label in enumerate(all_ids):
    print(str(label))
    plt.annotate(label, (x[n], y[n]))
plt.show()
Appendix D: Python script for image scraping

# I. Necessary Imports
import matplotlib.pyplot as plt
from PIL import Image
from pylab import figure

file = "path"
im = Image.open(file, 'r')
width, height = im.size

# II. Scraping
pix_val = list(im.getdata())
len = len(pix_val)
pointsx = []
pointsy = []
totalx = []
for x in range(len):
    if pix_val[x][0] >= 200 and pix_val[x][1] <= 20 and pix_val[x][2] <= 20:
        totalx.append(x)
y_pixel = x/width
x_pixel = (x/width - int(x/width))*width
pointsx.append(int(round(x_pixel)))
pointsy.append(int(round(y_pixel)))

# III. Reduce coordinates
pointsx2 = []
i = 3
while i < (len(pointsx)-7):
    i += 1
    add = 0
    for x in range(1,7):
        if totalx[i+x] < (totalx[i]+7):
            add = x
    totalx2.append(totalx[i])
i = i + add

pointsx2 = []
pointsy2 = []
for x in range(len(totalx2)):
    y_pixel = totalx2[x]/width
    x_pixel = (totalx2[x]/width - int(totalx2[x]/width))*width
    pointsx2.append(int(round(x_pixel)))
    pointsy2.append(int(round(y_pixel)))

pointsx3 = []
pointsy3 = []
nietdoen = []
bereik = 2
for x in range(1, len(pointsx2)):
    if x not in [nietdoen[c] for c in range(len(nietdoen))]:
        for b in range(1, len(pointsx2)):
if pointsx2[b] >= pointsx2[x]-bereik and pointsx2[b] <= pointsx2[x]+bereik and pointsy2[b] >= pointsy2[x]-bereik and pointsy2[b] <= pointsy2[x]+bereik:
    nietdoen.append(b)
pointsx3.append(pointsx2[x])
pointsy3.append(pointsy2[x])

print(len(pointsx3))
fig = plt.scatter(pointsx3, pointsy3, marker='.')