3 Analysing Knowledge Capture Mechanisms: Methods and a Stylised Bioventure Case Study

Abstract
Knowledge transfer between science and technology has been studied at micro and macro-levels of analysis. This has contributed to the understanding of the mechanisms and drivers, but actual transfer mechanisms and processes, be they through codified or tacit sources, have very rarely been mapped and measured to completeness and, to a large extent, remain a black box. We develop a novel method for mapping science-technology flows and introduce ‘concept clusters’ as an instrument to do so. Using patent and publication data, we quantitatively and visually demonstrate the flows of knowledge between academia and industry. We examine the roles of exogenous and endogenous knowledge sources, and of co-inventors and co-authors in the application of university-generated knowledge. When applied to a stylised case study, we show that the method is able to trace the linkages between base knowledge and skill sets and their application to a technology, which in some instances span over twenty-five years.

3.1 Introduction
Knowledge transfer between universities and firms has become increasingly institutionalised (Geuna & Muscio, 2009) as universities look for novel, more insightful, ways to enhance their economic and societal value through new technology spin-offs or start-ups (Audretsch et al., 2005; Tijssen, R.J.W., 2006). Much of the previous literature has focused on the facilitating actions and conditions for knowledge transfer such as scientific publications, conferences, informal interactions, collaborative and contract research, IP licensing, personnel exchanges and hiring - each with varying significance for industry (Ponomariov & Boardman, 2012).

A major challenge to evaluating these knowledge transfer routes and mechanisms is uncovering meaningful linkages between technological outputs and scientific inputs. Knowledge transfer occurs most often at both the codified and tacit level, and the transfer processes and motivations within academic research versus those in industry settings are complex and evolving. However, what is not discussed in detail in the existing literature is the demarcation and measurement of the knowledge that is transferred (Bozeman, 2000). This is of utmost importance because the facilitation of transfer has been investigated but the question of whether knowledge has been transferred can only be answered by (a) being able to demarcate the object of transfer, and (b) measuring its point of inception, evolutionary path and eventual application.

---

1 This chapter has been published as Gurney, T. et al. (2012). Knowledge Capture Mechanisms in Bioventure Corporations. Proceedings of Science and Technology Indicators (STI), Montreal and has been submitted to The Journal of Informetrics.
Specific quanta of knowledge evolve along developmental paths, shaped not only by the scientific and technological developments of the laboratory in which it was conceived, but also by the further learning and skill sets of the scientists and inventors involved. By exploring the routes of codified knowledge transfer from inception to exploitation, we can begin to understand the processes and mechanisms of knowledge transfer. These include interactive knowledge production, the role a scientist’s skill set plays, the effect of a scientist’s peers - be they in the university or in the lab - and the transformative nature of science itself.

There have been substantive efforts to examine the facilitation processes and end-utilisation in an isolated sense, analysing each step in the overall development process of a technology. However, a novel methodological approach is necessary to address the question of whether the whole transfer process has occurred. As mentioned previously, this requires a method to both demarcate and track specific quanta of knowledge. Doing this grants us a clear view on the effectiveness of the facilitating conditions.

In this paper we use and adapt available tools and models, and integrate those with newly developed tools to provide a complete picture of knowledge transfer, from start to finish. This paper starts with a discussion of the role of the scientist/entrepreneur, and that of his surroundings, in developing the necessary skill sets and knowledge for eventual transfer and application in industry. We then apply these insights to our methodology, which is described in detail. A crucial step in the methodology is the introduction of the idea of ‘concept clusters’, which refers to a small, cognitively cohesive agglomeration of scientific peer-reviewed publications. As an illustration, we briefly apply the methodology to a case study. In the conclusions, we summarise the potential benefits, open methodological issues, and routes for further research.

3.2 Conceptual framework

The codification of knowledge takes two primary forms: patents and scientific publications. The use of patents as indicators was pioneered by Schmookler (1966), followed by many applications (such as Schmoch (1993) and Fleming (2001)). However, many aspects of their indicator-orientated uses do have drawbacks (Pavitt, 1988). For example, not all innovations are patented (Arundel, 2001; Arundel & Kabla, 1998), with many innovations kept under a veil of secrecy (Brouwer & Kleinknecht, 1999), leading to underestimation of innovative potential or capacity. Analyses using patent indicators are typically based on metadata found in patents. Title words, abstract words and keywords (Courtial et al., 1993; Engelsman & van Raan, 1994), patent classifications (Leydesdorff, 2008; Tijssen, R.J.W. & Van Raan, 1994), and patent/non-patent citations (Karki, 1997; Meyer, M.S., 2001) have all been used extensively. Many patent databases exist from which we extract the metadata used in analyses, each with their own idiosyncratic advantages and disadvantages. These include disclosure requirements of prior art (‘duty of disclosure’): the USPTO requires an exhaustive list but the EPO requires a minimal listing. Differences also stem from the databases themselves, in terms of their formatting, whilst others relate to the practices of applying for patents through different national or supranational patenting offices. Despite the stated shortcomings, patents can be used for mapping knowledge transfer in a large part of the knowledge-intensive economy because patent documents are highly detailed descriptions of the processes, applications and necessary information required for a technology. Citations within a patent document, either to other patent documents or scientific
literature, add to this wealth of data. Patent documents encompass a wide range of technological fields and the major patenting offices (such as the USPTO or EPO) cover patent data from all countries (Tijssen, 2001).

Publications serve as the primary indicators for the defining characteristics and development of science. They are the most visible outcome of scientific endeavours, with an extensive range of indicators and methodologies developed. The analysis of publications shares a number of analytical approaches with patent analyses, such as word mapping (Callon et al., 1991) and citation analysis (Garfield & Welljams-Dorof, 1992; White & McCain, 1998). Using co-occurrences of combinations of words and cited references in publications is also becoming a common technique (Braam et al., 1991; van den Besselaar & Heimeriks, 2006).

The act of publishing itself is subject to a complex system of social and scientific norms, practices and reward systems (Merton, 1957). Publishing behaviours and patterns of scientists are governed in large part by these norms and practices, as well as by serendipity. The development of a university scientist’s profile and portfolio are the result of search strategies (Horlings & Gurney, 2012) employed by the scientist. University-based scientists publish primarily to extend their professional and intellectual prowess, and regular publishing is considered a requirement. Industry-based scientists are governed by similar constraints, and the firm benefits from publishing too - by becoming intimately involved with the basic science behind the technologies (Rosenberg, 1990), and their publications serve as a signal of their capabilities to the outside world (Hicks, 1995).

The conditions required for facilitating the development and transfer of knowledge depends heavily on the recipient knowledge platform. Knowledge assets (Nonaka, 1994), sector roles (Baba et al., 2009) and science-push and demand-pull concepts (Langrish et al., 1972), are all factors in a knowledge base’s receptivity. In this manner - external knowledge sources, taking into account demand and current capabilities, are readily absorbed and entrained into stock knowledge bases and practices. This receptivity is known as ‘absorptive capacity’ (Cohen & Levinthal, 1990) and can best be described as “[t]he ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends is critical to its innovative capabilities,” (p.128). The individuals involved are at the heart of this, with the absorptive capacity of a firm tied to its constituent individuals’ absorptive capacity, i.e. the right personnel are in place to take advantage of incoming information. As Cohen and Levinthal (1990) state, “Beyond diverse knowledge structures, the sort of knowledge that individuals should possess to enhance organizational absorptive capacity is also important. Critical knowledge does not simply include substantive, technical knowledge; it also includes awareness of where useful complementary expertise resides within and outside the organization” (p.133).

The use of patent and publication data, in the context of absorptive capacity, allows us to map knowledge inputs and outputs, and consequently illuminate the mechanisms at work. The aim of this study is to provide a map of the cognitive route between the scientific origins and the technological output, with specific focus on the knowledge capture mechanisms operating. To this end, we have developed a method that shows:
1. **How the scientific background of the patent corpus links to the scientific output of the inventor.**
   A patented product is the result of accretion over time of the research results, practices, skill sets and processes of the inventors involved. In the patent documents one can identify references to the underlying science (cited publications) and technology (cited patents) that were instrumental towards the development of the new (patented or non-patented) technology. Linking the patent corpus and publication output allows us to determine the background or necessary scientific requirements for the technologies.

2. **How the collaborative research environment of the researcher/inventor contributes to the development of the underlying science and to the technology developed.**
   Academic and industrial collaboration is common in high-technology fields. Much of science is the result of collaborative efforts between researchers, where resources can be pooled and task allocation increases efficiency. As such, any contributions from a researcher's network will be visible in any publication authorship list or patent inventor list.

3. **What other knowledge is needed by the researcher/inventor for the development of a technology, and how this is appropriated.**
   Scientists must incorporate new results and skills from previous research done by others, to improve upon and modify their own intellectual prowess and breadth of skills. Their individual absorptive capacity of the individual is measured by their entrance into, and adoption and integration of, new fields cited by the technologies they work in.

By mapping these three aspects of the knowledge stream, we can clarify several of the mechanisms through which knowledge capture is supported: (1) the researcher/inventor's own research; (2) the researcher/inventor's collaboration network; (3) the researcher/inventor's knowledge uptake process.

In order to map aspect 1, we have developed an approach based on the overlap in content of the non-patent literature references (NPLRs) found in the patent applications, and the publication corpus of the inventor. An individual publishes in multiple streams of research, with the streams being composed of publications highly similar to each other, which can be determined algorithmically. The similarity between the researcher/inventor's publications and the NPLRs can be calculated so that the NPLRs are co-located together with the research streams of the individual.

Comparing the underlying total knowledge and skill set required to develop the technology with the knowledge and skills of the individual researcher/inventor shows the contribution of the latter to the technologies. The contributions of an individual's co-inventors and co-authors can be similarly constructed allowing us to map aspect 2. Finally, in order to map aspect 3, a more refined approach is required. An individual's research streams may be broad in topic and time, and general statements can be made regarding the relevance and importance of an individual's knowledge and skills to a company's technologies. In order to examine the specific scientific fields that the technology draws upon (as defined by the NPLR and the fields from which they originate), we have developed a method focusing on the specific scientific concepts and methods necessary for the technologies described in the patent documents. By identifying the specific concepts utilised in the technologies and the point in time that the researcher/inventor develops or integrates them into his or her knowledge base, we are able to view from where, and from
what original form he or she derived new knowledge assets. This method utilises concept clusters, which will be defined and operationalised in detail in the next section. Concept clusters are used to map aspect 3 and to examine the detailed concepts and methods in aspects 1 and 2.

### 3.2.1 Concept clusters

A broad description of an individual’s knowledge and skill sets may be derived through examination of the titles used, references cited, keywords used (and more) in their publication corpus. Adding the NPLRs of the patent applications to the individual’s publication corpus allows us to discern which aspects of an individual’s corpus are similar to the NPLR. To discern general research themes within the combined corpus, we utilise the Louvain clustering method (Blondel et al., 2008) which optimises the modularity of a network, i.e. the actual distribution of edges between nodes versus a random distribution, to identify macro-clusters in the network. The metadata occurrences in each cluster are then examined to identify the general themes. To identify specific topics, each macro-cluster is isolated and the same clustering algorithm is applied to produce micro-clusters. These micro-clusters constitute the immediate environment of the NPLRs. Depending on the variety of subjects in the publication corpus, macro-clusters can range in size from 10 to 100+ publications whereas each micro-cluster is typically no larger than 10 publications.

We refer to these micro-clusters or immediate environments as ‘concept clusters’. The publications cited by the patent applications (NPLRs) make up the nucleus of the concept cluster and each concept cluster contains at least one NPLR. Surrounding this nucleus are the publications most similar in terms of title word and cited reference combinations (van den Besselaar & Heimeriks, 2006), and the borders of each concept cluster are algorithmically delineated into communities (Blondel et al., 2008). A concept cluster contains, in varying proportions, publications authored by the researcher/inventor (which the patent applications may or may not cite), and publications written by others that are cited by the patent application. The specific composition of a concept cluster describes the knowledge utilised in the patent application, in terms of the knowledge base and skills internal or external to the researcher/inventor.

<table>
<thead>
<tr>
<th>Cited by patent</th>
<th>Publication authored by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Inventor</td>
</tr>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>No</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 1  Concept cluster composition
Table 1 illustrates the possible publication origin types - A, B and C - in a concept cluster. For each concept cluster, a mix of publication type can result. If the concept cluster contains:

1. Type A publications - this indicates direct contributions by the inventor to the required concepts and skill sets. The research and concepts contained within the publication are either necessary for, or directly related to, the development of the technology.
2. Type B publications - we assume that some knowledge is outside the expertise of the inventor.
3. Type C publications - whilst the inventor is not cited directly, his or her publications are highly similar to publications that are being cited. The inventor's skill sets and background knowledge are similar to the NPLR.

As defined previously, absorptive capacity is the ability to recognise, assimilate and integrate new knowledge, and apply it in a novel manner. The absorptive capacity of an individual who is an inventor of the technologies can be determined by analysing the similarity and presence (or lack thereof) of their contributions to the concept clusters. The greater the number of occurrences of their own publications that are similar to the NPLR, the higher the enabling potential for absorptive capacity.

The timing of an inventor's publishing entry into a concept cluster is important. If publications by an inventor appear (Type A or C) first, followed by NPLRs (Type B), we conclude that the inventor has already previously developed the skill sets and knowledge required for the technologies. If NPLRs appear first and are then followed by the inventor's publications, we conclude the inventor previously did not have the skills or knowledge necessary but has had to address this. How soon an inventor publishes after recognising the NPLR indicates the perceived importance of that knowledge to those technologies, and to the inventor's own knowledge stock. This increases the similarity between the NPLR and the IA's publications, and consequently the absorptive capacity.

Using this approach, we can map aspect 3 from the previous section, and add it to the first two. We determine whether an inventor is a leader in the production of the knowledge required for their technologies, or a follower. If a follower, does the inventor incorporate the necessary knowledge and skill sets into their portfolio early, demonstrating a high level of absorptive capacity? If necessary knowledge and skills lie outside the portfolio of the inventor, do his collaborators provide any of the knowledge or skills?

3.3 Previous work

Previous studies typically utilise text-mining approaches or citation matching to provide a linkage between patents and publications. Text-mining approaches generally involve methodologies that identify topical clusters in patents and publications using words (title, abstract, or full text) and link the two corpora together through the similarities between the topical clusters. Mogoutov et al. (2008) use a combinatorial approach to map innovation in the biomedical field of microarrays. Relevant concepts are extracted from multiple data sets, namely those of publications, patents, and research project data. A matching algorithm links the data sets through their shared concepts. They specifically try to avoid using pre-determined topic areas or research areas, to allow some qualitative room for interpretation after the matching has been completed. They successfully demonstrate a link between, and within, scientific fields through shared concepts.
Magerman et al. (2010) provide a very thorough review of the state of the art of the text-mining approach. In addition, their study tests the effectiveness of distance measures when linking patents and publications via text mining. With only 30 patents and 437 publications, Magerman et al. use a smaller data set than what is typically encountered. These are notable figures, because commonly-used similarity distance measures rely on large data sets to provide high-quality matching outcomes. The authors acknowledge this and conclude that the overall number of records would likely increase the chance of linking patents and publications.

Text mining can rely on an abundance of methods, which are highly variable and customisable. However, some limitations of text mining also become apparent. The different vocabularies employed between patents and publications pose a threat to accurate matching. The size of the sample may result in misleading or inaccurate matching options. A further limitation is one of a changing vocabulary over time within a field of science. In publishing, the audience and indexer effects (Leydesdorff, 1989; Whittaker, 1989) may lead to fewer and fewer matches between publications and patents further apart in time. Text mining is typically a resource-intensive approach, and requires extreme care due to the complex nature of linguistic behaviours and anomalies.

Citation matching is easier as it involves extracting the bibliographic non-patent literature references (B-NPLRs) from the patent documents and finding the corresponding twin in whichever publication database one uses. Unfortunately, the requirements for including citations (patent and non-patent) in patent applications vary drastically between patenting offices, making the duty of disclosure a prime example relevant to this study. The move to include in-text non-patent literature references (IT-NPLRs) is a recent development as the availability of extraction tools for full-text documents has increased. A study by Tamada (2006) addresses the issue of IT-NPLRs, focusing specifically on Japanese patent documents. They argue that as there is no requirement by the Japanese Patent Office to include front-page references, patent output indicators that utilise only B-NPLR may miss relevant scientific references. To counter this, they use references found in the text of the documents to successfully identify under-reported scientific fields cited by patent applications. They conclude that the inclusion of both in-text and bibliographic citations enriched their data sets and provided balance between objective and strategic referencing of literature in patent applications.

Meyer (2002) examined the use of citations in patent and publication-centred studies. He formed a typology of the most frequently used approaches, such as patent citation analyses, industrial scientific activities, and university and academic patenting. His critiques of the techniques essentially point to the misuse of analytical tools and methods from one field to another. He notes that techniques that use these approaches do not take certain fundamental basic characteristics of patenting and publishing into account. For instance: firstly, different fields show a different propensity to publish; secondly, citations can be negative or positive; and thirdly, publishing is not the only output of the laboratory. In terms of patenting, similar problem characteristics should be taken into account, such as: the patenting propensity varies across industries, not all inventions are patented, and a significant proportion of patents are strategic, designed to block innovation by a competitor. Meyer may have examined these aspects over ten years ago, but the principles remain valid today when discussing methodologies using citation behaviours of publications and patents. Regarding NPLRs in patents, the relative abundance of
references to scientific literature versus non-scientific literature is an indicator of the quality (Branstetter, 2005) and proximity to science (Callaert et al., 2006) of the patent application. What is generally understood and accepted is that placing citations to scientific literature in patent documents indicates a cognitive link to, or awareness of, the related scientific concepts (Tijssen, R.J.W., 2001).

3.4 Data and Method
The methodology we developed consists of various steps. The first step is to select the inventor/researchers that play a crucial role in the relevant knowledge transfer case study. The second step is data collection of the papers and patents of these individuals (4.1). Then we do publication clustering (4.2) and patent application clustering (4.3). The next step is to link the patent applications and publication clusters, using a specific visualisation tool (4.4). After having described the method, we demonstrate it in section five: proof of concept.

3.4.1 Data
There are many patent databases around, all with their own idiosyncrasies, some of which stem from the databases themselves whilst others relate to the practices of various national or supranational patenting offices. In our study we use the PatSTAT database prepared and developed by the EPO, as it aggregates various other databases, and is considered one of the most extensive patent databases. For our publication data, we use the Thomson Reuters’ Web of Science (WoS) as our primary source of publication data, supplemented by CV data from the scientists involved. The sources and types of data come from:

1. Patent data - we extracted all patent applications with the selected inventors each listed as an applicant from the EPO PatSTAT database along with all other inventors’ data; this also included all patent applications with the firm under study listed as an inventor or assignee, and the selected inventors as assignees.

2. Publication data\(^2\) - we extracted from WoS all publications with the inventors’ firm listed as an institution; and all publications with the selected inventor listed as author.

These base data were parsed using SAINT\(^3\)(2009) and managed in a relational database. Further data were collected from the patents. More specifically, these were (and where they were found):

1. Bibliographic NPLRs (B-NPLRs) - these are citations included primarily by the examiner and added as front-page references.

2. In-text non-patent literature references (IT-NPLRs) - citations to publications visible in the body of the patent. These IT-NPLRs were automatically extracted from the full-text versions of the patent documents by custom software.

\(^2\) English language only
\(^3\) SAINT (Science-system Assessment Integrated Network Toolkit – a Rathenau Instituut open-source software suite designed to parse, clean and organise bibliometric data to be used later in relational database software such as MS Access and MySQL.)
All patent applications were then grouped according to their first filing, with the priority patent application representing the collective. Single-priority based families are collections of patent applications that claim a specific application as the first or priority application. The priority patent is included in the collection (Martinez, 2010). This is done to account for variations in NPLR reporting and inclusion across different patenting offices. A second reasoning is that any derivative applications are close extensions of the priority patent, thus one could expect the NPLRs from the collective to extend to the other applications in the group. Further references in this paper to these patent collectives use the term ‘priority patent’ to mean the priority patent application representing the collective.

The NPLRs were then normalised for search of their twin in WoS. If there were no NPLRs linked to any given priority patent, the NPLRs of derived citing patent applications (i.e. applications citing the original application as priority) were included. Both NPLR sets were parsed and, as far as possible, their WoS publication equivalents found. A manual check was performed to see if the retrieved documents matched the original NPLR. If any discrepancies in metadata did not allow for a proper match, the affected records were not utilised in any further analysis. The verified documents were then parsed and processed together with the inventor’s publications to create a master publication database and the origins of each document were recorded.

3.4.2 Publication similarity and concept clusters
Publications are clustered by their shared combinations of title words and cited (van den Besselaar & Heimeriks, 2006). The degree of similarity is calculated using the Jaccard similarity coefficient. Clusters of publications were automatically assigned by a community detection algorithm (Blondel et al., 2008) grouping publications based on their degree centrality and relative weights of edges between nodes. These clusters are referenced further as ‘research streams’. Each research stream is then isolated and the community detection algorithm of Blondel et al. is run on the individual streams resulting in the concept clusters.

3.4.3 Patent clustering
Patent applications are grouped by INPADOC family ID and the NPLRs of the INPADOC families within concept clusters are noted. The Jaccard similarity coefficient is calculated between INPADOC families using the shared concept clusters in which their NPLRs occur, and the community detection algorithm of Blondel et al. is used to designate INPADOC clusters.

3.4.4 Visualising patents and publications
We have developed a method (Horlings & Gurney, 2012) that allows the specific research trails that an individual has developed to be visualised in a uniquely clear manner. We have built upon this method by adding patent applications whose individual researchers are listed as an inventor to their corpus of publications. The thematic and knowledge base aspects of the patents and publications are linked, not through direct citations by patents to the publications, but through

---

shared thematic or topical research areas of the cited NPLRs and the inventors’ corpora of publications. Even if the patent document does not cite the individual’s publication corpus directly, other cited literature may (or may not) cluster within the inventor’s areas of expertise. This approach results in a tangible, visible, shared knowledge base between the patent and the publication.

We arranged the patent applications and research streams along two axes: time on the x axis, research streams and patents on the y axis. Longitude is defined as \[(\text{year of publication}) - (\text{year of first documented publication in the data set})/(\text{range in years}) \times 360\] - 180. Please note that the small clusters of papers visible within each research stream represent the annual production within that stream, and not the concept clusters - the latter contain papers published over multiple years. Latitude is defined as \[(\text{stream number})/\text{(total number of streams}) \times 180\] - 90. The nodes were positioned with the GeoLayout in Gephi (2009), using an equirectangular projection.

3.5 Proof of concept
In this section we apply the methodology to a stylised case study - stylised, as we are primarily interested in demonstrating that the methodology is able to map the knowledge streams as well as the mechanisms underlying these streams. This is based on a real case study that we aim to analyse in depth elsewhere (Gurney et al., 2013). The essential mechanics of the methodology are discussed here, with additional detail provided in the in-depth study. Images here are stylised, with data utilised to illustrate the necessary components.

3.5.1 Case study selection
Our case study involves a prominent biotechnology researcher who is strongly involved in cancer therapeutics at the firm he founded in 2001 and the university at which he is a professor. (The individual, firm and university shall further be referred to as IA, FA and UA respectively.) IA maintains direct links between his research at UA and research conducted at FA. This enables us to draw upon his extensive publishing history as well as his numerous patenting activities at both university and firm.

3.5.2 Data summary
Table 2 shows a summary of the various data collected. Patents cover the period 2000-2008\(^5\), and the publications cover all publications in the categories defined in the previous section up to 2011. The large number of patent applications (242) may not be typical of most companies in this field. The breadth of the patent applications, as exemplified by the number of INPADOC families (90) is also large. IA is a prolific author with, in 2011, 931 publications to his name. This is an exceptionally high number and we assume many of these publications are purely the result of him being head of a large institute in which his name appears as author as a matter of seniority.

---

5 Patent applications up to 2008 were chosen as there is considered to be a delay in the completeness of patent data in PatSTAT. 2008 was chosen as the last year as we could be more certain that all possible patent data was included.
Table 2  Summary of collected patent and publication data of IA

<table>
<thead>
<tr>
<th>Data</th>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patent Applications</td>
<td>Patent applications with FA listed as assignee and IA as inventor (2000-2008)</td>
<td>242 (115 priority patent applications)</td>
</tr>
<tr>
<td></td>
<td>INPADOC families</td>
<td>90</td>
</tr>
<tr>
<td>Publications</td>
<td>IA publications retrieved from WoS</td>
<td>931 (786 pre-2009)</td>
</tr>
</tbody>
</table>

1. **Mapping the links of the scientific background of the patent corpus to the scientific output of the inventor.**

Figure 1 is a stylised image showing portions of the total corpus of IA publications and the NPLRs of the patent applications differentiated into research streams. Noted in Figure 1 are the publications authored by IA and/or publications cited by the patent applications over time. The visualisation was constructed as detailed in section 4.4. Research stream **Rb** contains a considerable number of NPLRs authored by IA as seen by the black nodes. Streams **Ra** and **Rc** contain NPLRs not authored by IA, publications authored by IA but not cited and very few NPLRs authored by IA co-located in the same stream. Each stream may contain a mixture of publication types (A, B or C), and the proportional presence of IA's publications (cited or not) in the stream indicates the proximity of IA's research to the research cited by the patent applications.

**Figure 1** Stylised image of patent applications and publications - authored by IA and/or cited by the patent applications over time

---

Note: For descriptions of type A,B,C nodes - see table 1
In Table 3, the NPLR distribution in the patent applications is noted. Most of the NPLRs come from within the text of the patent documents. 65 NPLRs are found in both the text and bibliography of the applications. IA’s publications make up 10% of the NPLRs, with a proportionally larger number being cited in the bibliography.

**Table 3. Summary of NPLR**

<table>
<thead>
<tr>
<th>Data</th>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPLRs</td>
<td>Unique NPLRs found (and matched to WoS) in all</td>
<td>525+</td>
</tr>
<tr>
<td></td>
<td>patent applications</td>
<td>(2037 total)</td>
</tr>
<tr>
<td></td>
<td>In B-NPLR only</td>
<td>147+</td>
</tr>
<tr>
<td></td>
<td>In IT-NPLR only</td>
<td>313+</td>
</tr>
<tr>
<td></td>
<td>In both B-NPLR &amp; IT-NPLR</td>
<td>65+</td>
</tr>
<tr>
<td>NPLR citations</td>
<td>Count of IA’s publications cited in NPLR (Note: % = x/a)</td>
<td>55 (10%)</td>
</tr>
<tr>
<td></td>
<td>In B-NPLR only (% = x/b)</td>
<td>18 (12%)</td>
</tr>
<tr>
<td></td>
<td>In IT-NPLR only (% = x/c)</td>
<td>19 (6%)</td>
</tr>
<tr>
<td></td>
<td>In both B-NPLR &amp; IT-NPLR (% = x/d)</td>
<td>18 (27%)</td>
</tr>
</tbody>
</table>

We can conclude from Figure 1 and Table 3 that aspects of research conducted by IA are relevant and necessary to the technologies represented by the patent applications. Some aspects are not within IA’s expertise, such as those in research stream Rc. IA’s research is cited in many instances in both the text of the document and the bibliography. Research that is cited but not authored by IA is often very similar to IA’s publications, as seen in stream Ra.

2. **Mapping contributions of the inventor’s research collaborations in the patent and publication data.**

Figure 2 is a stylised image of portions of the total corpus of IA publications and NPLRs. The presence of co-inventors as authors of publications is noted. In research stream Rc, there are a number of NPLRs authored by IA’s co-inventors but do not feature IA as an author. In stream Rd there is a high proportion of NPLRs and non-NPLRs authored by both IA and his co-inventors. Stream Re contains many publications that are authored by both IA and his co-inventors but not cited by any of the patent applications. All three streams contain many papers by IA which were not co-authored by the co-inventors, and these papers are partly NPLRs, and partly non-NPLRs. The streams also contain NPLRs not authored by IA or his co-inventors.
Figure 2  Co-inventors of IA and their presence in NPLR and publications

Table 4 summarises the presence of co-inventors’ publications as NPLRs in the whole corpus. Only 9 publications written by IA’s co-inventors (without IA as author) are cited by the patent applications in the NPLRs (8 are shown in stream Rc in Figure 2, the last is located in another stream) and most are IT-NPLR (in-text). The number of co-inventors’ publications cited by patent applications (excluding publications co-authored by IA) is far lower than the number of IA’s publications cited by the applications. IA’s co-inventors appear as inventors without IA on 30 patent applications.

Table 4  Summary of co-inventors’ publication and patent application contributions

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-inventors</td>
<td>Publications Cited in NPLR (excluding publications co-authored by IA)</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>In B-NPLR only (%=x/b)</td>
<td>2 (1.5%)</td>
</tr>
<tr>
<td></td>
<td>In IT-NPLR only (%=x/c)</td>
<td>5 (1.5%)</td>
</tr>
<tr>
<td></td>
<td>In both B-NPLR &amp; IT-NPLR (%=x/d)</td>
<td>2 (9%)</td>
</tr>
<tr>
<td>Patent applications</td>
<td>Publications not cited by patent applications but co-authored with IA</td>
<td>251</td>
</tr>
<tr>
<td>Patent applications</td>
<td>FA patent applications without IA as inventor</td>
<td>30 (12%)</td>
</tr>
</tbody>
</table>

Note: for b, c and d values see Table 3
3. **Mapping the inventor's level of adaptive knowledge use (absorptive capacity) necessary for the development of a technology.**

We map the concept clusters and their utilisation by patent applications over time, in order to demonstrate the absorptive capacity of IA in relation to the development of the technologies. To construct the concept clusters, each research stream in the total corpus is isolated. The community detection algorithm of Blondel (2008) is run over each isolated stream. Each resulting community or concept cluster contains a mixture of publications (Types A, B and/or C from Table 1), with at least one NPLR forming a nucleus. We grouped each patent application into INPADOC clusters, based on the similarities of the aggregated main group IPC codes of the patent applications' parent INPADOC families. The resulting INPADOC clusters represent the different technologies in which IA is involved, and their development over time.

Figures 3-5 show stylised representations of the appearance in time of the concept clusters containing NPLRs cited by the INPADOC clusters. The concept clusters represent not only the immediate knowledge environment of the NPLRs, but also the degree of similarity between the inherent skill sets of IA and the skill sets referenced by the patent applications in the NPLRs. This allows us to map the knowledge and skill sets that are necessary for the development of the technology (and therefore cited by the patents). IA's own publications may not be cited in many instances but are highly similar to the NPLRs and are co-located in the same concept cluster. This mapping strategy also shows the stage of the technologies’ development at which the skill sets and knowledge from outside IA’s expertise are cited.

Figure 3 includes concept clusters that contain only NPLRs not authored by IA (Type B from Table 1). The skill sets and knowledge base contained in these publications are considered to be outside the expertise and skill sets of IA as they do not contain any similar IA publications (in terms of title words or cited references).

**Figure 3** Concept clusters cited by INPADOC clusters containing only NPLRs not authored by IA (Type B publications)

Note: Concept clusters are labelled CC. Size of nodes indicates count of publications and count of INPADOC families. Thickness of edges indicates number of citing INPADOC families. Edge colour indicates at what phase in the age of the INPADOC cluster the concept is cited, grey=early, dashed=middle, black=late.
As is visible in Figure 3, the NPLRs located in the concept clusters are cited at different stages in the development of the technologies of INPADOC clusters 1-3. For example, CC b/4 (a concept cluster found in research stream Rb) is cited in the middle development period by INPADOC cluster 1, but at a later period by INPADOC cluster 3 whereas CC c/1 (a concept cluster found in research stream Rc) is cited early in the development of all the INPADOC clusters. CC a/1 from Ra, is cited in the middle development phase by only INPADOC cluster 1.

**Figure 4** Concept clusters cited by INPADOC clusters containing NPLRs not authored by IA (Type B) & IA publications not cited by patent applications (Type C)

Figure 4 shows concept clusters containing a mixture of NPLRs not authored by IA (Type B) and publications authored by IA but not cited by the patent applications (Type C). In the highlighted area 1 are two concept clusters (CC c/3 and c/4), both from research stream Rc. These are cited in the early phases of development of INPADOC clusters 1, 2 and to a lesser extent 3, and CC c/3 is cited in the middle phase of development by INPADOC cluster 2. IA’s publications appear early in CC c/3 but late in CC c/4. This implies that the knowledge and skill sets of CC c/3 are necessary from an early period and IA is publishing in this concept cluster also at an early stage, whereas IA only starts publishing in CC c/4 at a late stage in the concept cluster’s lifespan. In highlighted area 2, concept clusters CC a/5 and a/6 from research stream Ra are cited in the early to middle development phase of the technologies. In CC a/5, IA’s publications are found early in the lifespan of the concept cluster, but in CC a/6, his publications only appear in the middle stages of the concept cluster’s lifespan.

Relating this to IA’s absorptive capacity: IA has already published similar publications taking into account the macro-clusters (considering CC c/3 & c/4 and CC a/5 & a/6 are from the same respective research streams) and highly similar publications considering each concept cluster.
separately. For the concept clusters in which IA’s publications appear only in the middle or late phases, we can conclude that IA has recognised the importance of the research being cited by the patent applications and, demonstrating a degree of absorptive capacity, begins to populate the concept clusters with his own publications (not cited by patents).

**Figure 5** Concept clusters cited by INPADOC clusters containing NPLR authored by IA (Type A), NPLR not authored by IA (Type B) and IA-authored publications not cited by patent applications (Type C)

Figure 5 demonstrates more clearly IA’s direct knowledge and skill contributions to the technologies, as is shown by citations to publications authored by IA in the concept clusters. As also found in Figure 4, the concept clusters are cited at different stages of development. In Figure 5, however, the concept clusters also contain NPLRs authored by IA. In many cases, concept clusters containing NPLRs authored by IA are cited during the early stages of the span of a concept cluster, and others in the middle stages. Many of these come from the same research streams (for example CC d/4, d/1 & d/6 are all from research stream Rd) and are cited by all three technology groups. In the case of CC d/4, there are some transitive similarities to d/1 and d/6, and IA only begins publishing in the middle stages of that concept cluster’s lifespan.

This once again demonstrates the absorptive capacity of IA because the research by IA that is necessary to the technologies is often cited early. Some of his research is cited later on. These publications appear in the middle stages of the concept clusters, but are eventually cited. In other words, aspects of his overall research have been necessary for the technologies and in areas in which he was not cited and/or active, he began research that eventually led to it being incorporated and cited.
Summarising the results seen in Figures 3-5, the scientific publications cited by the patent applications in the INPADOC clusters stem from three sources. These sources include 1) publications cited by the patent applications but not authored by IA, 2) publications authored by IA but not cited by the patent applications, and 3) publications by IA that are cited by the patent applications. The composition of the concept clusters and the period of citing by the INPADOC clusters indicate the relevance of the concepts to the technologies at different times. The entry of IA publications into these concept clusters indicates IA having a degree of similarity in knowledge and skill sets to the cited publications. The period of entry by IA's publications indicate the adoption of these skills and knowledge by IA. As per the definition by Cohen & Levinthal (1990), absorptive capacity is the ability “… to recognize the value of new, external information, assimilate it, and apply it to commercial ends [and this] is critical to its innovative capabilities.” In this sense, the entry by IA publications into the NPLRs at varying time periods is indicative of the ability of IA to recognise, assimilate and integrate external knowledge into his own skill sets and knowledge base, and these integrations consequently become visible in the scientific background of the patent applications. We also observed IA starting new research in order to improve his absorptive capacity where he seemed to have gaps in knowledge and skills.

3.6 Summary and conclusion

The diverse characteristics of knowledge production, incorporation and dissemination relating to product development result in a complex model of knowledge capture. Previous methods used to investigate knowledge transfer have focused on the facilitating conditions with little concern paid to whether there is any actual knowledge transfer. In this paper, we have explained and developed a method to demarcate and track knowledge transfer. We have done so by combining and modifying existing techniques and supplementing them with new methodological tools. The resulting method allows us to address the more complex aspects of knowledge capture mechanisms - as illustrated with a stylised start-up or spin-off case.

With our methods for data processing, clustering and visualisation, we can demonstrate the thematic and theoretical links of the inventor’s patent output and the inventor’s knowledge base and skill sets, as represented by their publication output. This marks a departure from the problems of previous methodologies that relied on individual-specific direct citations from patent applications to literature in order to determine the theoretical influences of an individual or a field in general (Meyer, 2002).

Our method allows for a close examination of the multidirectional aspects of linkages between science and technology. This provides a quantitative measure of how effectively, and from where exactly, an idea generated in academia makes its way into an industrial application or, conversely, how skills and knowledge developed in application may be followed by new lines of research generating new scientific knowledge and skills.

Our approach to determining the absorptive capacity of an individual allows us to evaluate the utilisation of scientific knowledge by individuals and their eventual application in technology. The methodology accounts for the influence of co-inventors on the combination of knowledge required for technological output. This allows us to determine the degree and field of contribution from the respective inventors in terms of the base knowledge required for the development of a technology.
We explained and demonstrated the methodology using a stylised case study in which one individual is responsible for much of the spin-off firm's growth and success and bridges both the academic and industrial aspects of knowledge transfer. His research is both fundamental (at his university setting) and applied (in the firms' appropriation and implementation). The researcher-inventor bridges the research environments, facilitating knowledge transfer and skills development between them. In further research we aim to investigate the same processes if there are more star or bridge scientists in one firm, and the effect of their overall contributions.

Our method is not without its shortcomings. Using in-text citations of patent documents requires a significant amount of cleaning due to the differing citation reporting behaviours and requirements across patenting offices. The correct assignment of authors and inventors to publications and patents requires a significant amount of disambiguation. Initially this was done by algorithmic means (Gurney et al., 2012) and then checked manually, which was a time-consuming process. The inclusion of all the NPLRs in the patent applications introduces a level of uncertainty because some of the NPLRs are not directly related to the technologies. Many NPLRs are very general in nature and address only the fundamental background of the technologies. These NPLRs are difficult to identify without comprehensive expert examination of the patents but are still included.

This new method of mapping science and technology output and the relationships between them deepens our understanding of the level of contributions made by individuals and firms, and also by specific institutional policies and models. If a firm or an individual carries out research within a specific research climate or environment, by utilising this methodology one would expect to see the overall publishing and patenting activity, and the links between the two, to vary according to the research climate or environment. This therefore enables empirical investigation of the influence of the environment on knowledge transfer and absorptive capacity.

3.7 References


Martinez, C. (2010). *Insight into different types of patent families: OECD.*


Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics, 26*(1), 193-211.


