Novel perspectives on university-industry knowledge transfer: A structural assessment and text mining application

Woltmann, Sabrina

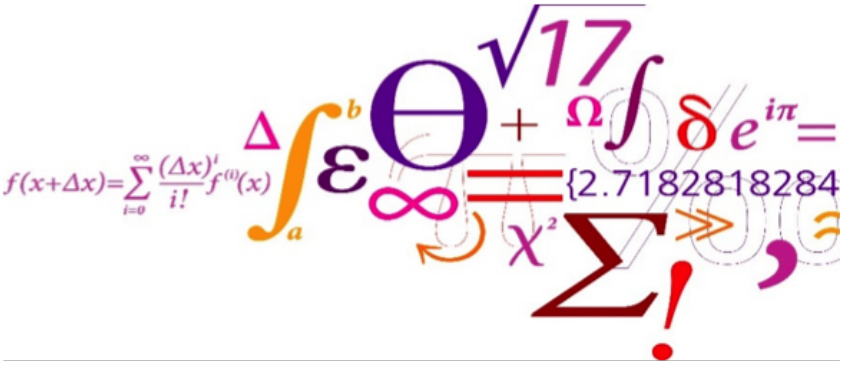
Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

NOVEL PERSPECTIVES ON UNIVERSITY–INDUSTRY KNOWLEDGE TRANSFER: A STRUCTURAL ASSESSMENT AND TEXT MINING APPLICATION

SABRINA LARISSA WOLTMANN

PhD thesis
Kgs. Lyngby, November 2018



Technology Innovation and Management (TIM), Department of Management Engineering, Technical University of Denmark
Diplomvej 375, 2800 Kgs. Lyngby, Denmark

SUPERVISORS

Associate Professor Lars Alkærsig (main supervisor)
Technology Innovation and Management (TIM)
Department of Management Engineering
Technical University of Denmark
Denmark

Associate Professor Melanie Kreye (co-supervisor)
Engineering Systems
Department of Management Engineering
Technical University of Denmark
Denmark

Associate Professor Carina Lomberg (co-supervisor)
Technology Innovation and Management (TIM)
Department of Management Engineering
Technical University of Denmark
Denmark

ASSESSMENT COMMITTEE

Professor Jason Li-Ying
Technology Innovation and Management (TIM)
Department of Management Engineering
Technical University of Denmark
Denmark

Research Fellow Pablo D'Este
Spanish Council for Scientific Research
and Polytechnic University of Valencia
Spain

Professor Magnus Gulbrandsen
Centre for Technology, Innovation and Culture
Faculty of Social Sciences
University of Oslo
Norway

This thesis was written with LaTeX using classicthesis v4.6.

I want to thank Lars Alkærsig for choosing me for this interesting PhD project and, more importantly, for giving me the opportunity to work on a challenging project in the three years of my PhD. I value highly that the project allowed me to expand my knowledge from social sciences towards statistics, machine learning and other technical methods. I am very thankful that you gave me the opportunity to develop the project in the direction I wanted to and that you gave me the freedom to make my own mistakes My PhD project was fully funded by DTU and I would like to thank DTU for the opportunity to visit many conferences, summer schools and interesting courses.

I would also like to mention my changing supervisor team: Line Clemmensen from DTU Compute, who showed me the direction and relevant courses to set a solid foundation for the technical requirements of the PhD project. Carina Lomberg and Melanie Kreye, who volunteered and helped me through the last critical months of thesis writing. As part of my PhD, I stayed at Georgia Institute of Technology from August 2017 until December 2017. Here I want to thank Professor John Walsh for his efforts to enable the stay, his supervision and engagement. I would also like to thank the other PhD students in TIM at DTU MAN and at DTU Compute for being supportive and inspiring.

I am most thankful for the support of the three main individuals in my life that made this PhD project possible: Maya K. Gussmann, Marc J. Silberberger and Nicolas Obriot. I have to thank all three of you for your patience, your (technical) help, your support, and, in particular, your encouragement. This would not have been possible without your knowledge and help all along the three years. In particular supporting me with understanding the statistics, programming and mathematical expressions was crucial for the success of this project! Maya's knowledge about R and Latex saved me months of trial and error, thank you. I am sure you will miss my strange emails with cryptic questions about error messages. Marc's knowledge about R and article writing that helped me to describe my technical progress in scientific terms. He also helped to decode R from a beginner's perspective making it less frustrating.

Finally, Nicolas, who dedicated many hours to help me understand basics in programming, APIs, SQL and so much more. He

helped me through the tough data collections and challenges. I am truly grateful that you three were there for me until the end.

And finally, there's my family and all my other friends that supported me during these exciting and challenging three years- thank you for being there, listening to me (sometimes for hours) and for making it an amazing journey.

"An expert is a person who has made all the mistakes that can be made in a very narrow field."

— Niels Bohr

SUMMARY

Universities face increasing demands for active dissemination of their re- search results and are expected to contribute to knowledge development in their socioeconomic environment. Universities are expected to be key drivers promoting economic development and innovation. Consequently knowledge dissemination, as a crucial aspect for industrial development and innovation, is politically highly desired and became a focus area for public funding of university research. Scholars, policy makers and practitioners picked up on this increasing demand for understanding university contributions and investigate collaboration and knowledge transfer between universities and industry.

However, some elements in the interaction between universities and industry that contribute to its effectiveness still remain largely unknown. Questions remain regarding especially the knowledge transfer channels and measurements of successful knowledge dissemination. The overarching aim of this PhD project is to identify novel potential measures for university-industry knowledge transfer through specifically chosen and adapted computational methods, hereby contributing to the understanding of university research knowledge transfer.

First, publication data from a single technical university's publication database were analysed regarding their distributions and ratios in different dimensions, such as publication types, research fields, etc. Additionally, coverage of long-standing established publication databases was taken into consideration. The results showed that the traditional databases have skewed coverage and novel or less traditional outcomes of research (output that is not a journal article or a book chapter) often might be significantly underrepresented. It shows that additional data can increase the insights into university research in certain aspects significantly.

In the second part of the PhD project, a novel approach for detecting knowledge transfer was developed and used to trace the content from university research in companies. Text mining applications were used to detect content from academic publica-

tion abstracts on company websites. The findings show that the detection of common content between universities and industry via text mining applications is possible and beneficial.

In the final part of the PhD project the methods are applied to investigate the impact of Open Access publications on knowledge transfer. Using the text mining methods, I examine the differences between subscription-based and Open Access publications, assuming that the accessibility of a written item implies a different performance in terms of knowledge transfer. Here the results show that for this specific measure Open Access publishing makes a difference in terms of university-industry knowledge transfer. Given the contemporary positive assumptions regarding Open Access publications, the differences appear less pronounced than expected.

Overall this thesis finds that novel computational methods can be used to detect knowledge transfer, but that further advancements in terms of technical tools and methods are needed to improve their performance and feasibility.

RESUME - DANSK

Universiteterne står over for et stigende krav om aktiv formidling af forskningsresultater og de forventes at bidrage til vidensudvikling i deres socioøkonomiske område. Universiteterne forventes også at gå forrest, når det gælder fremme af økonomisk udvikling og innovation. Derfor er vidensformidling af afgørende politisk betydning for industriel udvikling og innovation, og det er blevet et fokusområde for offentlig finansiering af universitetsforskning. Lærere, politikere og praktikere forholder sig til denne stigende efterspørgsel efter en dybere forståelse af universiteternes bidrag ved at undersøgte samarbejde og vidensoverførsel mellem universiteter og industrien. Imidlertid er nogen af de aspekter, der bidrager til en effektiv interaktion mellem universiteterne og industrien, endnu relativt ukendte. Der er stadig ubesvarede spørgsmål med hensyn til kanalerne for og målingen af succesfuld vidensoverførsel.

Det overordnede mål med dette ph.d.-projekt er at identificere nye potentielle metoder til måling af vidensoverførsel fra universiteter til industrien gennem specifikt udvalgte computerbaserede metoder. I den første del af dette projekt blev publikationsdata fra et enkelt teknisk universitets publikationsdatabase analyseret med hensyn til udbredelse og i forhold til publikationstyper, forskningsområder mv. Der blev taget højde for dækning af længe etablerede publikationsdatabaser. Resultaterne viste, at de traditionelle databaser har en skæv dækning og nye eller mindre traditionelle forskningsresultater (output, der ikke er en artikel eller et bogkapitel) er ofte særdeles underrepræsenteret. Dette viser, at yderligere data kan betyde en signifikant større indsigt i universitetsforskning inden for visse aspekter.

I anden del af ph.d.-projektet blev der udviklet en ny tilgang til detektering af vidensoverførsel. Denne nye tilgang blev anvendt til at spore resultatet af universitetsforskning i virksomheder. Tekst mining applikationer blev brugt til at registrere indholdet fra akademiske publikationer på virksomhedernes hjemmesider. Resultaterne viser, at det er muligt og udbytterigt at finde fælles

indhold mellem universiteter og industrien ved hjælp af tekst mining.

I den sidste del af ph.d.-projektet anvendes metoderne til at undersøge effekten af Open Access-publikationer på vidensdeling. Ved hjælp af tekst mining undersøger jeg forskellene mellem abonnementsbaserede publikationer og Open Access publikationer, idet jeg antager at tilgængeligheden af en publikation har betydning for vidensoverførsel. Her viser resultaterne, at Open Access gør en forskel. I betragtning af de positive antagelser omkring Open Access-publikationer synes forskellene dog mindre udtalte end forventet.

Samlet set konkluderer denne afhandling, at computerbaserede metoder kan bruges til at registrere vidensoverførsel, men at yderligere fremskridt med hensyn til tekniske værktøjer og metoder er nødvendige for at forbedre deres ydeevne og effektivitet.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 PURPOSE & OUTLINE

Universities face increasing demands for active dissemination of their research results and are expected to contribute to knowledge development in their socioeconomic environment. Universities are expected to be key drivers, promoting economic development and innovation. Consequently, knowledge dissemination, as a crucial aspect for industrial development and innovation, is politically highly desired and became a focus area for public funding of university research. Scholars, policy and practitioners picked up on this increasing demand for understanding university contributions and investigate collaboration and Knowledge Transfer between universities and industry. However, some elements in the interaction between universities and industry that contribute to its effectiveness still remain largely unknown. Some questions remain, especially regarding the Knowledge Transfer channels and measurements of successful knowledge dissemination activities and strategies. The overarching aim of this PhD project is identifying novel potential measures for university-industry Knowledge Transfer through specifically chosen and adapted computational methods. This thesis aims to increase hereby the understanding about university research Knowledge Transfer. The goal was first to investigate potential data structures of university research outcome, and second how the successful transfer can be identified in a novel manner. To address the purpose of the thesis, three research objectives were developed and investigated in three single studies. The first part of this PhD project focused on an analysis of university research data. Data from a single technical university's database were analyzed, including their distributions and ratios in different dimensions, such as publication types, research fields etc. Additionally, coverage of long standing established publications databases was taken into consideration. The results showed that the traditional databases have skewed coverage and novel or less traditional

outcomes of research (output that is not a journal article or a book chapter) might often be underrepresented. It shows that additional data can increase the insights into university research in certain aspects significantly.

In the second part of the PhD project, a novel approach for detecting Knowledge Transfer was developed and used to trace the content from university research in companies. I used text mining applications for the detection and as main data sources academic publication abstracts and company websites content. The findings show that the detection of common content between universities and industry via text mining applications is possible and beneficial. In the final part of the PhD project, some approaches are taken to investigate the impact of Open Access publications on Knowledge Transfer. Using the text mining methods, I examine the differences between subscription based publications assuming that the accessibility of a written item implies a different performance in terms of Knowledge Transfer. Here, the results show that for this specific measure, Open Access publishing makes a difference in terms of university-industry Knowledge Transfer. Given the contemporary notions on Open Access, the findings are not as significant as I expected.

Overall this thesis finds that novel computational methods can be used to detect Knowledge Transfer, but that further advancements in terms of technical tools and methods will help to improve their performance and usability.

The purpose of this PhD study is to improve the understanding of university-industry Knowledge Transfer, by investigating the internal knowledge structures and measuring successful dissemination. The goal is to identify the contribution universities, in their role as public knowledge centers, provide for innovation and problem solving in their (economic) environment. This thesis aims to develop novel insights into research knowledge structures and the measurement of university-industry Knowledge Transfer. The PhD thesis itself is based on three manuscripts that are executed as independent studies. It has the following structure:

THE INTRODUCTION gives an overview about practical relevance, motivation, and the political and socioeconomic background and the literature basis for the thesis.

CHAPTER 3 focuses on the relevant academic literature, summarizes contemporary empirical studies and introduces the conceptual framework of the thesis.

CHAPTER 4 describes the research strategy and methodology for the thesis.

MANUSCRIPT I describes the university as knowledge base and provides in-depth insights about the structures of an internal knowledge system, including its composition over time. It sets the base for understanding university knowledge structures.

CHAPTER 6 introduces the computational (text mining) methods used and tested in the Manuscript II and Manuscript III.

CHAPTER 7 describes the data samples and collection which used in the two final manuscripts. It provides a large amount of detail regarding the data collection process and relevant technical decisions.

MANUSCRIPT II introduces the main methods adapted and tested for the detection of university-industry Knowledge Trans-

fer.

MANUSCRIPT III uses the methods to identify potential differences between Open Access and subscription based publications in terms of university-industry Knowledge Transfer.

THE DISCUSSION AND CONCLUSION (CHAPTER 10) summarizes the main findings of the individual manuscripts and clarifies their interrelation. It describes the theoretical, empirical and policy implications. It concludes on the research aim of the thesis and gives a brief summary about the limitations and future research possibilities.

Introduction & Background

Gives an insight about the scope and theoretical framing of the thesis project

Objective 1: *"Identify structures and changes to university research knowledge output"*

RO1

Manuscript I:
"University Research Knowledge Structures and Their Development Over Time"

Objective 2a and 2b:
"Develop/adapt computational methods to detect university-industry knowledge transfer"
"Evaluate the potential of these methods in terms of their potential."

RO2a, RO2b

Manuscript II:
"Tracing university-industry knowledge transfer through a text mining approach"

Objective 3: *"Use the methods to investigate potentially relevant dimensions of university-industry knowledge transfer"*

RQ3

Manuscript III:
"Open Access' influence on University-Industry Knowledge Transfer"

Discussion & Conclusion

Provides a summary of the findings and concludes with the main insights the project provides

Figure 1: Overview over the objectives, the manuscripts included in the thesis, and the related research questions.

2 INTRODUCTION

Universities typically are seen as major contributors to knowledge creation, innovation and scientific understanding in societies (Bercovitz and M. Feldman, 2006). Their historic legacy as knowledge centers, as hubs for exceptionally intellectually (scientifically) gifted individuals is well established and explains the special role of universities in society. Universities have made a societal contribution for centuries through the education of students and independent research. For many years, the primary role of universities was to foster innovation through scientific breakthroughs and cutting edge technologies (Scott, 2006). That is, they were increasingly expected to contribute to their (national) socioeconomic environment through knowledge development (Jongbloed et al., 2008). Nowadays, universities are facing additional demands for active dissemination of their research results (B. R. Martin, 2011). This change in the public's perception has led to a third university mission in addition to its research and teaching roles i.e. the active dissemination of their research outcomes (Gulbrandsen and Slipersaeter, 2007). This mission, to disseminate knowledge is described as the third Mission (e.g. (Etzkowitz et al., 2000; Laredo, 2007). Universities are being expected to disseminate novel research-generated knowledge that can be utilized and adapted by the private sector. In turn, its adaption is supposed to contribute to economic development and innovation in national contexts. The (knowledge) contribution of universities is often described as university impact (D'Este et al., 2018). University impact has been discussed and examined in the academic literature and is an objective of politicians and other stakeholders (Rasmussen et al., 2006) to legitimate extensive public funding for university research. Over past decades, evidence based justification became essential because public funding is a limited resource subject to claims from many different public institutions (Geuna, 2001) and politicians have an incentive to

allocate resources to the best benefit for society. As a result, universities and stakeholders are seeking to provide evidence of the (societal) outcomes of their contributions. However, university impact is complex, and not easily understood or explained, and is difficult to measure empirically. So far, its underlying mechanisms and implications are not fully understood (Bozeman, 2000; Santoro and Bierly, 2006). In particular, the transfer of knowledge from universities is a key aspect of assessment of the impact of universities. Although prior research provides several well developed indicator frames, there are some significant gaps related to the detection and measurement of this Knowledge Transfer (Agrawal, 2001). Given the priority and relevance of this, understanding the elements underlying university impact based on evidence of Knowledge Transfer is crucial. New evidence will not only provide justification for public spending but also lead to insights into the potential of (and potential improvements to) university impact through Knowledge Transfer. This makes this topic relevant for policy makers as well as universities, practitioners and the academic community in general (Pesole, Nepelski et al., 2016).

Based on this strong need for a better understanding and better measurement of university impact (Cheah, 2016; Gherardini and Nucciotti, 2017), this thesis aims to enhance our understanding of the contribution made by universities through Knowledge Transfer. Specifically, the aim is to identify and measure Knowledge Transfer. To achieve this, I investigate whether current understanding of university-industry Knowledge Transfer can be enhanced by the application of novel computational methods.

My overall aim can be split down into four smaller research objectives (ROs) which are addressed in three different studies, which are part of this thesis:

Research Objective 1 (RO1): *"Identify structures and changes to university research knowledge output."*

Research Objective 2a (RO2a): *"Develop/adapt computational methods to identify university-industry Knowledge Transfer."*

Research Objective 2b (RO2b): *"Evaluate these methods in terms of their potential."*

Research Objective 3 (RO3): *"Use the methods to investigate potentially relevant dimensions of university-industry Knowledge Transfer."*

Knowledge Transfer can be seen as an input-output activity involving two parties both of which must be understood and measured. Hence, the first objective (RO1) involves in-depth understanding of the university research. In this context, potential data sources and indicator frames are examined to understand the structure of and changes to university knowledge output. In the case of RO2a and RO2b, it is necessary to choose, understand, adapt and test novel computational methods to identify university-industry Knowledge Transfer. In the case of RO3, I investigate certain conditions on the university side which might facilitate university-industry Knowledge Transfer.

Methodologically, the aim is to provide a first proof of concept of the newly adapted methods, and an assessment of future potential. The empirical scope is a technical university in Scandinavia and technical sciences. This university has extensive links to industry and seems therefore particularly appropriate for this investigation. The sources of data include several publicly available databases such as the university's own publication database, patent databases (e.g. PATSTAT) and the online publication database Scopus. Additional data were gathered from online web crawling. These data combined constitute novel data which are important in the context of new perspectives and measurements (cf. chapter 7).

The investigation should provide a) an improved understanding of established data sources and their usefulness to legitimize further empirical research b) identification of potential novel methods to trace Knowledge Transfer; c) increased understanding of features relevant to university knowledge dissemination. Although I cannot claim to be inventing a new holistic measure, I contribute by establishing an additional measure based partly on new data sources and not yet used technical methods which helps to fill the current (academic) knowledge gap. Overall, this work contributes to the understanding of universities as knowledge centers. It offers insights into text based measurement possibilities and provides evidence showing whether less conventional approaches might in future be an adequate extension to present empirical methods. Finally, the findings should be

understood in the context of particular policies and university activities.

2.1 BACKGROUND AND POLICIES

To understand the scope of this thesis, I need to present some of the (national) legal and political realities in which universities act. The political paradigm shifts and changes to incentive structures in particular are relevant to this broader context. The European, Scandinavian and Danish contexts are discussed to provide a more structured and straightforward reading of the thesis.

2.1.1 IPR Regulations and Changes

University activities and strategies are restricted and driven by legal regulations, policy frameworks and funding opportunities (Munari et al., 2016; Siegel et al., 2007). Knowledge dissemination is crucial for industrial development and innovation (D'Este and Perkmann, 2011). It has become a political focus in the context of public research funding and is shaping the regulatory frameworks in many (national) contexts (Cattaneo et al., 2016; Hicks, 2012). As a result, many institutions and funding agencies are requiring submission of detailed information on university impact, and especially university-industry collaboration. The intention of legal amendments regarding university driven innovation is related to the same incentives and concepts as justification of public funding - namely, to increase the benefit to society of research outcomes. Making the knowledge accessible, usable and beneficial for the economy is seen as one way of achieving this target.

Given that universities need to acquire and secure public funding to ensure their working basis, Knowledge Transfer is extremely relevant to them. In fact, public funding is the main source of income for most universities (Auranen and Nieminen, 2010; Bentley et al., 2015). Not surprisingly, industry friendly activities and extensive collaboration strategies have become the declared goals of most universities today (Geuna and B. R. Martin, 2003).

Aligned to these objectives, the commercial value of research has received particular emphasis since the 1980s. Many policy adaptations and legal changes are focused on concrete research outcomes, and in particular intellectual property rights (IPR) regulation and scientific publications. An important change occurred in 1980 with the US Bayh-Dole Act. The Bayh-Dole Act was an attempt to remove potential obstacles to university-industry technology transfer through legislation (Mowery et al., 2002). The Act constitutes a uniform patent policy across all federal agencies; it removed previous restrictions on university licensing, and allowed universities to own the patents deriving from research funded by federal research grants. Before the Act, patenting rights were the property of federal governments. It was assumed that universities had few incentives to invest in commercial side of research if the federal government was the main beneficiary of those efforts. The Act was aimed at giving universities the flexibility to negotiate licensing agreements, and to give firms more incentive to engage in collaborations. "The framers of this legislation asserted that university ownership and management of intellectual property would accelerate the commercialization of new technologies and promote economic development and entrepreneurial activity." [p. 112] (Siegel et al., 2003). The overall intention was to facilitate and speed up utilization of publicly funded inventions (Kenney and Patton, 2009). The effectiveness of the Bayh-Dole Act is debated, and the empirical evidence is mixed. Some scholars consider it has improved the relationships among and exchanges among universities and industry practitioners (J. Thursby et al., 2001), while others attribute the increased patenting activity to external circumstances and see little evidence that the law has provided the promised benefits (J. Thursby and M. Thursby, 2003).

Despite lack of clear evidence, many European countries have followed the US example and adopted similar legislation to change ownership of IPRs to university inventions. However, these changes in Europe apply to a very different reality. While in the US federal government owned the inventions prior to the Act, countries such as Denmark and Germany previously granted professors the right to retain IPRs over their research findings according to the so called professor's privilege. In Denmark,

the notion of professor's privilege was established in 1955 but this changed in 2000 when Denmark became the first of several European countries to abolish professor's privilege by granting IPRs to the university. This shift has been shown often to be ineffective and insignificant with the vast majority of originally academic patents remaining the property of firms (Lissoni et al., 2009). Therefore, the legislative changes in Europe are considered misguided, and focused on the wrong aspects of university patenting activities: "European universities today and in the past (before the abolition of any privileges) add substantially to patenting, although they tend to leave most patents in the business partners' hands" (Lissoni et al., 2009)[p. 12]. Several studies show that in Europe compared to the USA academic patents are much less likely to be owned by universities, since European universities seldom have the same autonomy and administrative capacity as their US counterparts (Lissoni et al., 2008). As a result, these general changes continue to be contested and are not necessarily considered improvements to the situation of research commercialization, university-industry collaboration, or fostering of innovation. Whatever the outcome, the intentions behind the policy changes remain clear: to improve the use and commercial value of university research, increase exchanges with private industry and ensure disclosure of novel knowledge.

2.1.2 Publications and Open Access

Pure commercialization is by far not the only aspect of policies and legislations aimed at facilitating the utilization of publicly funded inventions. A more recent phenomenon is the focus on public availability of publicly funded research through the lifting of subscription-based restrictions on academic publications. Publications traditionally have been used as the main means of communication among researchers (Jokić et al., 2018), guaranteeing research quality and enabling quantitative assessment of university research performance (Hicks, 2012). Scientific publications have been described also as one of the most important sources of learning about public research (Picarra et al., 2015). Hence, the traditional subscription-based publications model is increasingly seen as (illegitimate) exclusion of the public through

excessive subscription fees which hamper scientific knowledge dissemination (Armstrong, 2015). Some scholars maintain that access of society to publicly funded research is a human right (Tennant et al., 2016). Based on these rather novel notions governments, policy makers, founding bodies and universities around the globe have adopted policies and agendas to facilitate a new publication strategy known as Open Access (School of Electronics and Computer Science at the University of Southampton England, 2018). The UK, for instance, has played a leading role in the establishment of an Open Access agenda and states 'Open access to research enables the prompt and widespread dissemination of research findings.' (Higher Education Funding Council for England, 2016)[p. 3], and some smaller countries such as Denmark have also adopted explicit Open Access policies. The notion of Open Access has been a heavily debated topic in the last few decades resulting in the first adoption of Open Access policy in 2012 by the several Danish founding bodies including: The Danish Council for Independent Research, the Danish National Research Foundation, the Danish Council for Strategic Research, the Danish National Advanced Technology Foundation and the Danish Council for Technology and Innovation (Danish Council for Independent Research, the Danish National Research Foundation et al., 2012). In 2018, the Danish Ministry for Science and Higher Education released 'Denmark's National Strategy for Open Access' and stated it to be '(...) a question of achieving the maximum effect from research' (Ministry of Higher Education, 2018)[p.1].

However, despite the emergence of Open Access policies and agendas, there remains a clear lack of consistent evidence based assessment since most studies of Open Access focus on the impact of Open Access publications on the academic community (Antelman, 2004) and neglected the implications for society and industry. The policy and legislative changes described above demonstrate the changing reality for universities. These changes have led to strategic adaptations by universities including novel strategic goals based on additional incentive structures which are shaping behaviors and decisions about research outcomes and dissemination. All of the policy changes and adaptations described point to the political and economical relevance of

university research, and underline the emphasis on knowledge dissemination from universities. However, they highlight the lack of evidence-based decision-making and the gap in our understanding of the key elements that might foster these goals. The contradictory evidence and highly diverse findings suggest that there continues to be a need to identify the relationships between university driven Knowledge Transfer and publication availability.

3 LITERATURE, CONCEPTS AND MODELS

The literature on which this thesis draws is comparatively broad and comprises assessments university research impact, knowledge and technology transfer, and considers also knowledge management and information systems and communication theory. The main focus is university-industry knowledge transfer assessment which provides an understanding of the role of universities as key drivers of the contemporary knowledge economy (Etzkowitz and Leydesdorff, 1995). It includes relevant concepts and definitions of knowledge and knowledge transfer. The integration of university-industry knowledge and technology transfer literature reveals the underlying mechanisms and drivers (D'Este and Parimal Patel, 2007; Etzkowitz et al., 2000). Additional perspectives are from the knowledge management (Liyanage et al., 2009) and information systems literature (Kuhlthau, 1991) and basic communication theory (Shannon, 1948). This increases the scope to an understanding of knowledge dissemination and transfer.

This chapter describes the literature, the underlying assumptions and the conceptual frames and builds the foundations for the frameworks employed in the accompanying manuscripts. It additionally identifies the gaps in the contemporary empirical literature and contemporary indicator based measurements of university-industry knowledge transfer.

3.1 UNIVERSITY IMPACT

Public-funded university research is regarded widely as playing a crucial role in economic development (Cheah, 2016) and is required to have positive implications for society (Bornmann, 2013). Universities are seen as knowledge centers and important parts of national innovation systems (Mowery and Sampat, 2005). This

(public) perception of universities has emerged over time and is driven by changes in public expectations towards universities as public institution. Their long established roles as teaching and research institutions are no longer the only focus with strategic dissemination of research outcomes that benefit society and industry as their *third Mission* (Laredo, 2007). This third Mission requires increasing links between universities, government and industry.

The main pillar of the third university Mission is the contribution of universities to the economy. It is linked closely to economic growth and R&D to provide novel knowledge and drive innovation. It is high on (national) political agendas. Conceptually the third Mission is embedded in the triple helix model (Etzkowitz and Leydesdorff, 1995), which describes and analyses the close ties between universities, government and industry, and assumes (mutually) beneficial interaction and exchange among these three components (Lengyel and Leydesdorff, 2011; Leydesdorff and Etzkowitz, 1998). The main contribution of universities in this perspective is seen as knowledge creation and dissemination of knowledge to industry (Etzkowitz, 2008), while government is seen mainly as shaping the framework for knowledge exchange and providing research funding. In line this view of Knowledge Transfer from universities, I examine two component of the system (H. W. Park et al., 2005). Several studies focus only on two components of the three-way relationship, here the most frequently studied components being universities and industry and the exchanges between them (Leydesdorff, 2012; Meyer et al., 2014). Hence, investigations of university-industry (in a double or triple helix format) collaborations constitute a large part of the literature on university research impact assessment (Kwon et al., 2012; Leydesdorff, 2012).

3.1.1 Knowledge and Knowledge Transfer

As the main component of the universities' contribution is dissemination and sharing of the knowledge created knowledge. Hence, *knowledge* is the key concept for this thesis. Knowledge is a relatively abstract concept and has been discussed by various academic disciplines (Zagzebski, 2017). It is conceptualized in

different but often complementary philosophical approaches dating back to Plato (Gettier, 1963). According to the Oxford Online dictionary, knowledge is "facts, information, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject" and for a single topic or field, knowledge is defined as "the sum of what is known" about the specific topic [1]. Knowledge can be distinguished from pure data or information by its requirement for human reflection on the information or experience. Hence, knowledge is content dependent and seen as a resource that is located in an individual or a collective (De Long and Fahey, 2000). It facilitates understanding of complex relationships and enhances decision-making capacity. In this regard, knowledge has different contexts and dimensions. The knowledge management literature splits knowledge into three main categories:, *Human Knowledge* which is closest to the Oxford Dictionary definition and which is constituted by individual skills and expertise; *Social Knowledge* which exists only at the group level; and *Structured Knowledge* which is embedded in systems, processes and routines and is explicit and rules-based (De Long and Fahey, 2000). For instance, Organizational Knowledge which can be seen as a particular sub-set of social knowledge, is defined as:

> 'individual knowledge paired with that of other individuals in an organization. (...) When individuals pool their knowledge within an organization, that knowledge can give the organization advantages over others in the same field.' [2]

An additional classification driven by other literature has been developed for academic contexts in the field of Sociology is *Scientific Knowledge* which is important in the context of university research. Science is seen as public not private knowledge and Scientific Knowledge and truths as dependent on the scientific community, Merton 1970 after Phillips 1974. This understanding and framework was developed over more than six decades ago to investigate the quality and epistemology of knowing and

1 https://en.oxforddictionaries.com/definition/knowledge
2 http://www.businessdictionary.com/definition/organizational-knowledge.html

knowledge which are an explicit focus in the field of the sociology of knowledge (Phillips, 1974). Scientific knowledge can be broken down further into two qualitative components (Polanyi, 1962): *tacit* and *explicit (codified) knowledge*. Tacit knowledge is described as "non-verbalised, intuitive and unarticulated knowledge" Polanyi 1962 after lyivange) [p. 119]. It represents a form of know-how that is developed by informally acquired behaviors and procedures (Howells, 2002) [p. 872]. This form of knowledge is closely tied to the individual and its experience and cannot be stored or written down in any formal form. It is described as certain underlying expert knowledge. "Explicit or codified knowledge involves know-how that is transmittable in formal, systematic language and does not require direct experience of the knowledge that is being acquired (...)" (Howells, 2002) [p. 872]. Generally, research knowledge involves both tacit and explicit knowledge. Explicit knowledge provides an understanding and needs often to be accompanied by tacit knowledge in order for it to be used. Some consider that explicit or codified knowledge is merely a reduction and conversion of scientific knowledge which has tacit and explicit aspects to the messages containing the key information (Partha and David, 1994). According to this understanding of explicit knowledge, it is closely related to (research) Knowledge Transfer[3], since tacit knowledge because it is based on learning through experience is hard to observe and even harder to transfer. The explicit aspects of research knowledge can be stored and written down to enable other individuals who lack this particular knowledge to understand, potentially use and enhance it. Explicit knowledge does not require any direct interaction for its transfer. In the case of the research objectives in this thesis only explicit knowledge is considered since it is clearly identifiable and quantifiable and potentially can be traced at different levels of the process. Therefore, explicit knowledge is the main dimension investigated in examining Knowledge Transfer.

Although *(university-industry) Knowledge Transfer* emerged more than thirty years ago as a separate research area, its theoret-

3 In this thesis the concepts of *knowledge* transfer and *technology* transfer are used interchangeably. However, I acknowledge that technology transfer in certain circumstances may be a narrower concept (Agrawal, 2001)

ical base remains fragmented. It has been suggested that this scientific domain lacks a proper independent theoretical base (Gherardini and Nucciotti, 2017). Moreover, Knowledge Transfer is also rarely defined clearly in the contemporary literature including the knowledge management literature (Liyanage et al., 2009). Hence, it is important to identify the specific aspects of Knowledge Transfer in terms of explicit Knowledge Transfer. According to Argote and Ingram, 2000 [p. 152] , Knowledge Transfer is "the process through which one unit (e.g., individual, group, or division) is affected by the experience of another". This rather vague definition can be sharpened using Liyanage et al.'s (2009) definition where

> '(...) a knowledge transfer process has two main components, i.e. the source or sender that shares the knowledge, and the receiver who acquires the knowledge'. [p. 123]

This definition includes the transfer mechanisms and outcomes. I also adopt the expanded notion that Knowledge Transfer is

> '(...) the conveyance of knowledge from one place, person or ownership to another. Successful knowledge transfer means that transfer results in successful creation and application of knowledge in organizations' [p. 122].

Therefore, it should be emphasized that Knowledge Transfer becomes evident only if it is measurable at the receiver's end which requires its utilization or display by the knowledge receiver.

In the case of Knowledge Transfer between universities and industry there are some particularities that affect its mechanisms and success. From a firm perspective, knowledge acquisition is the basis of competitive advantage in terms of innovation and R&D (Argote and Ingram, 2000). However, universities are evaluated according to their contribution to the private economy and hence have an interest in sharing their research findings (B. R. Martin, 2011). The special university-industry relationship is characterized further by several unique features: First,

universities do not compete for segments in the private market and, hence, do (mostly) not compete directly with companies; Second, universities do not have a strategic interest in protecting their knowledge, rather they want to disseminate it (Laursen and Salter, 2014); Third, companies can outsource (usually to gain financial advantages) research areas that are not among their core competencies since this may be more efficient; and, Fourth, firms can rely on research based education for their employees if they engage in university collaborations. Accordingly, the parties are mutually dependent in order to achieve the maximum of knowledge creation and innovation.

3.2 CONCEPTUAL FRAMEWORK

A dedicated strand of literature on (university-industry) Knowledge Transfer has emerged and is the interdisciplinary basis for conceptual frames and extensive empirical work (Bozeman, 2000; Ramos-Vielba et al., 2009). To ensure that the underlying transfer process some particular features are considered when constructing a conceptual framework I reviewed the literature on Knowledge Transfer strategies (Rossi and Rosli, 2015), channels and mechanisms (Bekkers and Bodas Freitas, 2008), firms' learning and adaptation capacities (Bishop et al., 2011), and individuals' behavior and decision making (Perkmann et al., 2013). This literature includes the fields of information systems, university-industry collaboration (D'Este and Parimal Patel, 2007), communication theory and knowledge management.

The four components I identified to build a conceptual frame are:

1. The knowledge creator and sender, the university;
2. The knowledge receiver and user (the firm and/or the individual industry researcher)
3. The mode of Knowledge Transfer (medium and channels);
4. The transfer outcome or demonstrated use of the knowledge.

3.2.1 Knowledge Creator and Sender

There is a large literature on university-industry Knowledge Transfer that deals mainly with the sender and creator of knowledge: universities (Feller, 1990; Perkmann and Walsh, 2009; Salter and B. Martin, 2001). Some studies distinguish here between basic and applied research. These categories of research science are often used to refer to the potential (direct) applicability or usability of the research for industry (Bentley et al., 2015). Although most university research is viewed as relevant for future innovation, basic research is often perceived to have a longer time lag and to require further development for it to be useful for industry (Pavitt, 1991). University research can contribute in different forms and be used for problem solving and problem identification or the generation of novel ideas (Cohen et al., 2002).

As mentioned one of the main aspects of university knowledge creation is that it needs to have a codified or explicit output in order to be understood. Hence, it has to be captured to enable others to reuse and expand on it. At this level, the also implications of Knowledge Transfer activities and collaborative efforts are seen as the main features of universities. The actual generation of research and knowledge is often not investigated explicitly. I also need to consider the individual academic researcher who is seen as the initiator and collaborator and has particular motives and agendas that shape the exchange of knowledge and its success (Perkmann et al., 2013). Intentions, motivations and limitations are often examined in great detail (Perkmann and Walsh, 2009; Siegel et al., 2004). The knowledge generated can be manifested in different university outcomes which results in outputs such as academic publications, technical reports, presentations, consulting, etc. Overall, universities research outputs are considered to be their main commodity which is disseminated within and beyond academia. These outputs can be seen as inputs to university-industry interaction.

3.2.2 Knowledge Receiver- Companies and Industrial Researcher

Industry or the knowledge receiver can be understood at two different levels: organizational (i.e. company), and individual (i.e. individual researcher in the firm). The first receiver component, the company is a key factor in evolutionary economics. Evolutionary economics (Dosi and Nelson, 1994) and, in the same frame, the concept of national innovation systems (Nelson, 1992) state that companies need to research, adapt, innovate and select to retain and enhance their competitive advantage. Hence, companies have a high level of interest in innovating, learning and adapting knowledge to maintain competitiveness (Santoro and Bierly, 2006). Thus, knowledge is a crucial resource for most companies. However, they need also to protect their inventions and knowledge from competitors (Laursen and Salter, 2014). Accordingly, they need innovation, adaptation and protection strategies. At the same time it is becoming increasingly important to leverage external knowledge sources to reduce own research expenditure, and to draw on external knowledge and expertise which could lead to breakthrough inventions in the market (Brostroem, 2012; Roper et al., 2017).

However, companies differ in their features and resources and this affects their ability to leverage external knowledge sources. The potential to integrate (external) knowledge and innovation is called absorptive capacity which is defined as the ability to understand and identify new external knowledge, integrate it, and apply it for commercial purposes (Cohen and Levinthal, 1990) [p. 128]. Therefore, absorptive capacity determines the company's knowledge acquisition strategies and potential collaborations. It is relevant to collaboration and exchanges with universities and can be considered a determinant of the receipt and utilization of novel knowledge (Brostroem, 2012; Laursen and Salter, 2004). The concept of absorptive capacity explains that both the available knowledge and the receiver of the knowledge are determinants of the successful transfer (D'Este and Perkmann, 2011).

At the same time, the individual company researcher is important for the successful university-industry Knowledge Transfer. Individuals adopt certain behavioral patterns and strategies to

identify relevant knowledge and problem solutions. This is discussed in detail in the information systems literature (Wilson, 2000). Information seeking behavior which refers to individual strategies for identifying information is an important aspect in the information systems field (Kuhlthau, 1991). Drawing up on some notions about this behavior, which has well-established explanatory concepts, key aspects are including access to information and trust in information sources among others. Specific attention has been paid also to individuals in industry who collaborate directly with universities (Ankrah et al., 2013). This interpersonal dimension is a crucial aspect for collaboration and Knowledge Transfer. Generally, has been found that industry researchers have some features in common with academic researchers. However, they are influenced by the organizational structure in which they are embedded and the time allocated to information seeking. Individuals' decisions have an important impact on the probability of successful Knowledge Transfer, knowledge identification and the ultimate exploitation of this knowledge. Availability, reliability and time efficiency are important in these decision making processes (Wilson, 2000). The education level of the researchers influences their knowledge about data sources, information systems and the academic literature, and hence will affect final knowledge adaptation. The company can be seen as the meta-level end user while the industry researcher searches for information, evaluates its content and quality and then uses or discards it.

3.2.3 Knowledge Transfer Mechanisms

The literature on university-industry Knowledge Transfer focuses on features that allow novel knowledge to diffuse beyond academia. The mechanisms of Knowledge Transfer between universities and industry have been studied in depth (Bekkers and Bodas Freitas, 2008; Brennenraedts et al., 2006), and especially university strategies and activities aimed at knowledge dissemination (Drucker and Goldstein, 2007; Huggins and Johnston, 2009; Siegel et al., 2003). As already mentioned, universities have a strong interest in disseminating their knowledge and implementing concrete strategies and activities for its dissemination

and transfer since evaluations of their contribution to economic development and society are increasing (B. R. Martin, 2011).

The dissemination activities implemented are linked directly to the mechanisms of Knowledge Transfer which include different types of (formal) engagement, networking, contractual research agreements, etc. A great deal of research about mechanisms focuses on activities to foster dissemination in a commercial context; for instance, activities that are carried out by Technology Transfer Offices (TTOs) (Arundel, 2008), formal contractual collaborations with industry, and research spin-outs (Rothaermel et al., 2007). There are quite some studies that see the individual researcher as the central dissemination unit and focus on their motivation, characteristics and behavioral patterns. These individual features are often set in relation to the overall dissemination strategies and performance of universities, and are seen as contributing to or hampering knowledge exchange. These individual and institutional activities are described as forming the basis of dissemination, transfer and collaboration channels between universities and industry. To distinguish among particular interaction and dissemination modes, some Knowledge Transfer taxonomies have been developed. There are two main types of Knowledge Transfer: *formal* and *informal*. In informal Knowledge Transfer, property rights are secondary and the accompanying obligations are normative rather than legal (Link et al., 2007). Informal mechanisms and outcomes have received significantly less research attention and mostly involve non-contractual interactions between universities and the industry (Grimpe and Hussinger, 2013).

Formal transfer includes the final allocation of property rights and obligations, and is described by Arundel 2008 [p. 642] as Knowledge Transfer that eventually will "result in a legal instrumentality such as, for example, a patent, license or royalty agreement (...)". The mechanisms involved include: licensing contracts (J. Thursby et al., 2001), collaborative research projects, and consulting. Formal transfer since it is usually based on a formal agreement or contract is easier to detect and evaluate which has resulted in more empirical attention.

3.2.4 Sender-Receiver Interaction

The last receiver component is the final transfer outcome which refers to adaptation, use and acknowledgement of the knowledge by the receiver and should be observable or measurable (Brennenraedts et al., 2006). This final outcome is the actual proof for successful transfer which can be verified. The transfer modes described and their mechanisms are the bases for sender and receiver interactions and highly interlinked to the successful Knowledge Transfer. They need to be further considered in terms of final outcomes.

However, identifying the outcome and, hence, the successful Knowledge Transfer is difficult. In many empirical studies from the receiver's perspective there are no measures of Knowledge Transfer outcomes dissemination by the university is already considered evidence of successful Knowledge Transfer. Here knowledge is available and could be used by the receiver. However, its actual use or presence at the receiver end is often not measured (e.g. (Perkmann et al., 2013). This is problematic for research on Knowledge Transfer since it leaves its acquisition hypothetical.

Another limitation in the contemporary literature is that many assessments are based on the (subjective) opinion of the receiver. This means that the companies (representatives) are asked how much value or knowledge gain they estimate resulted from collaborations. There is often no additional objective validations of receivers' perceptions applied (Brostroem, 2012; Santoro and Bierly, 2006). It has been found that companies "(...) rate universities very low as information sources and potential partners, but their actual use and impact on firms is much higher" (Howells et al., 2012) [p. 706]. The result is misconceptions about the true value of university research impact, and especially university-industry Knowledge Transfer. There is a need to integrate the final outcomes of university-industry Knowledge Transfer, and develop objective outcome measures to generate a more holistic basis for empirical work.

3.3 FOR THE THESIS

The conceptual framework and the discussion above form the foundation for this thesis. In what follows, I describe how I use these different aspects to obtain a better understanding of the underlying mechanisms and parameters and the transfer of knowledge between universities and industry. The aspects of university-industry Knowledge Transfer are used mainly to understand the factors that are relevant to interpretation and analyses.

On the universities' side (as the knowledge creators and senders) the thesis considers in particular the aspects of research output and traditional dissemination strategies: academic publishing and, if adequate, university patenting. This captures the university knowledge base (see Manuscript I), and the features of knowledge dissemination (see Manuscript III). It also helps to identify and assess university output which in turn, represents the inputs to university-industry Knowledge Transfer. The behavior of individual academic researchers is not included explicitly or measured directly (e.g. publication decisions and strategies), since the thesis is not aimed at examining human behavior. However, the conceptual discussion is necessary to understand the underlying features.

On the industry side (as the knowledge seeker and user) I consider company characteristics and abilities which allow interpretation of the analytical findings. These are the notions underlying this work, and are relevant for generating the model but are not investigated empirically. This applies also to the individual behavior of industry researchers. In the case of transfer modes, I aim to apply no limitations to one or the other form of Knowledge Transfer so I include formal as well as informal transfers. It is important to include as many potential occurrences of Knowledge Transfer as possible. Therefore, decision inhibits that the results can be set in relation to a concrete Knowledge Transfer form. Nevertheless, we can identify the most important parameters influencing the outcome of university-Knowledge Transfer.

The above described components and interactions are the main underlying assumptions for this thesis. It adopts an input-output

based model which examines the input of university research and observable output at the firm end based on an understanding of the key components. So underlying notion presented above play a role in the understanding, but not every aspect is explicitly examined in the later empirical work. In (Manuscript III) these assumptions are further clarified and expanded to construct an expanded input-output model (for a better overview of the applied concepts and the content of the respective Manuscripts see Figure 2). Clearly, this rigid view of Knowledge Transfer from a sender-receiver perspective is an abstraction of the reality which in involves much more complex structures. For instance universities also benefit from collaborating with industry. However, this aspect is beyond the scope of the present study.

3.4 EMPIRICAL ASSESSMENTS AND MEASUREMENTS

Assessments of university performance in terms of their impact on society and the industry has led to the emergence of increasingly elaborate frameworks (B. R. Martin, 2011) encompassing several different features and aspects of Knowledge Transfer. As mentioned above, they are aimed at understanding the motives, activities, channels and facilitators of knowledge exchange (Franco and Haase, 2015; Rossi and Rosli, 2015). The empirical literature has proposed several measures of Knowledge Transfer. These empirical approaches are either qualitative studies (e. g. case studies) or quantitative such as econometric studies (Cheah, 2016). Qualitative and quantitative studies have been used to identify the underlying mechanisms and evaluate university performance in terms of Knowledge Transfer. The focus tends to be interactions and collaboration and their implications (Bruneel et al., 2010; D'Este and Parimal Patel, 2007). Qualitative approaches tend to investigate the motivations for university-industry collaboration and the channels of knowledge exchange. They focus often on single universities, projects or national contexts as case studies (Ankrah et al., 2013; Perkmann and Walsh, 2009; Rothaermel et al., 2007). Quantitative studies use different mechanisms and proxy-indicators to measure the frequency

and commercial implications of Knowledge Transfer (Henderson et al., 1998; Rossi and Rosli, 2015). For instance, Brennenraedts, et. al (2006) summarize the mechanisms including the indicators such as different transfer media used for knowledge dissemination. These indicators are not restricted to a single perspective or unit of investigation (see Table 1).

Table 1 presents the range of indicators and indicator groups some of which have been the subject of in depth investigation and especially indicators of publications, IPR and entrepreneurship have proven successful over the years (Agrawal, 2001; O'Shea et al., 2008; J. Thursby and M. Thursby, 2000). In particular, in quantitative works the commercially relevant indicators have dominated assessments of university contributions (Perkmann et al., 2013). For example, IPR related Knowledge Transfers has been used to assess the extent and value of knowledge dissemination for decades; these include indicators such as patents, co-patenting by companies and universities, licensing and royalties (Crespi et al., 2011; J. Thursby et al., 2001). These indicators are particularly useful because they identify university research output and provide norms and measures for commercially relevant inventions. However, not all research can be patented or licensed. Other studies examine output such as publications and co-publications involving companies and universities (Calvert and Pari Patel, 2003). These studies identify university output and the overlapping knowledge structures. However, all of these indicators has some limitations, and despite their extensive use do not capture the full extent of university-industry Knowledge Transfer(Agrawal, 2001).

3.4.1 Empirical Indicators for the Thesis

This chapter has pointed out that the measurement of university-industry Knowledge Transfer is subject to some limitations. Most of them are empirical in nature. In order make a theoretical contribution, I investigate the potential of novel methods (cf. RO2a and 2b). To do so, I adopt proven indicator frameworks as the foundation for my further empirical work. I develop some supplementary measures using the existing and partly novel indicators. In particular, since no existing measures can capture the

full picture of university-industry Knowledge Transfer, a combination of the established indicators and novel methods constitute a promising approach to capturing the contribution made by university research. Use of novel methods might allow for a better understanding and assessment. This thesis focuses mainly on academic publications, and in some cases, uses university patents as indicators of knowledge production. The importance of these indicators has been validated in previous studies. In addition, I include some unconventional indicators and data sources (see chapter7).

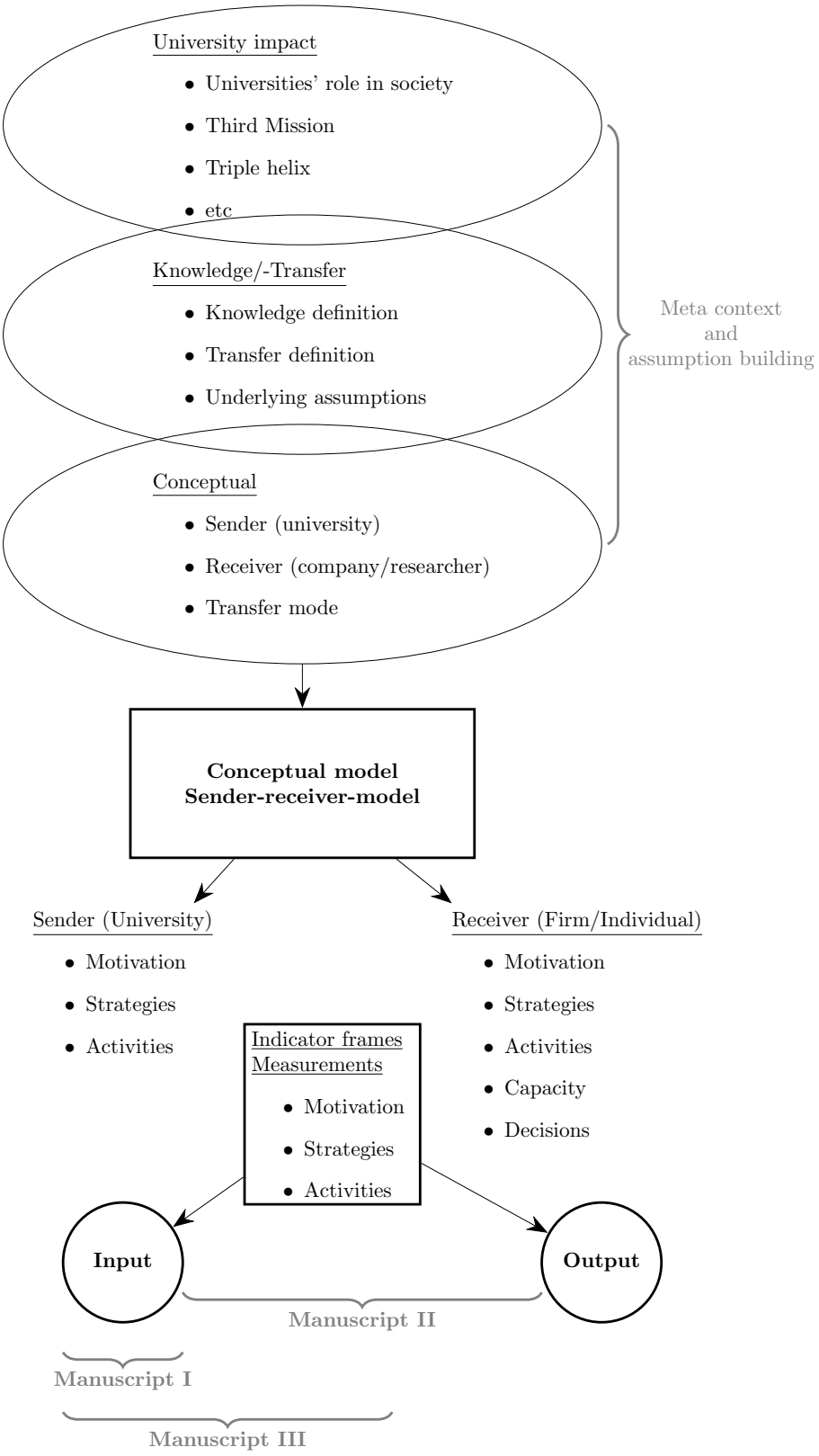


Figure 2: The combination of literature and concepts for the thesis and the respective Manuscripts

Transfer channel/ mechanism	**Related Indicators**
Publishing	Scientific publications Co-publications
Participation-informal networks	In conferences In boards of institutions ...
Mobility of people	Graduates Double appointments Mobility from institutes-industry
Informal contacts	based on friendship Alumni societies ...
Sharing of facilities	Shared laboratories Science parks ...
Cooperation in education	Contract education or training Influencing curriculum Providing scholarships Working students
Contract research	Contract-based research Contract-based consultancy
IPR	Patents (texts) Co-patenting Licenses of university-held Copyright
Spin-offs and entrepreneurship	Spin-offs Start ups Incubators at universities Stimulating entrepreneurship

Table 1: Different categories and forms of university-industry Knowledge Transfer and aligned indicators (adapted from Brennenraedts et. al 2006)

4 METHODOLOGY AND RESEARCH STRATEGY

Although the methodology and research strategy choices might seem obvious, it is important to understand some of the considerations and schemes underlying the thesis. This chapter should ensure common ground for the analysis, and especially interpretation of the results. Achievement of the overall goal involves several dimensions which require the application of both collective and individual strategies for the individual studies on which the depends. A broad goal of the thesis is to measure effects and causes by proposing new methods for their investigation. Rather than exploiting an exploratory approach, it adopts a deductive methodology (Robson and McCartan, 2016)[p.19]. Thus, the individual studies build on previously established knowledge in the form of theories to deduce research objectives which should allow the theories to be tested empirically in a particular context.

Embedded within the thesis aim are concepts that form the basis for the research objectives which are the final operationalizations of the theoretical assumptions. They constitute the foundation for further methodological decisions which shape the research strategy and data choices made for this thesis research. University-industry knowledge flows, and the impact of university research and Knowledge Transfer activities have been studied from a variety of perspectives. Quantitative approaches underlie most of studies and help to shed light on some of their key underlying principles. According to Creswell (1994), quantitative research is research that explain[s] phenomena by collecting numerical data that are analyzed using mathematically based methods (in particular statistics). Quantitative empirical research has proven to provide persuasive and conclusive results. Quantitative approaches are based on assumptions about the principles of objective research in which reality is a phenomenon that can be observed through the data collected and objective rather than subject measures.

However, it is important to note the interpretation and knowledge derived from particular results are to an extent socially constructed. Hence, this thesis research adopts not a positivist but rather a critical realist perspective (Sukamolson, 2007). The overall thesis aims to enable generalizations yielding to predictions and potential explanations which will feed back into the theoretical understandings about the structure of university knowledge systems and the transfer of research. This final inductive step constitutes the contribution to theory, and includes the main reflections and contextualization of the thesis findings. However, this does not deflect from the overall deductive nature of this research (Bryman, 2012; Perry and Jensen, 2001).

Quantitative research methods are appropriate for this thesis for two main reasons: First, the literature, research, and practice show that quantitative approaches are able to provide effective means of measurement (Agrawal and Henderson, 2002; Perkmann et al., 2011). They provide the foundations to verify theoretical and conceptual frameworks. Based on previous work, it would seem that a quantitative approach also facilitates phenomenon detection. Second, to address the research objectives and support or reject the related hypotheses, quantitative research combined with cautious interpretation could establish a new baseline for understanding the underlying reality. Quantitative measurements of the reality have advanced our understanding and allowed for a more general overview of Knowledge Transfer trends and patterns (Agrawal and Henderson, 2002).

The choices of methods and data sources were made to provide an adequate baseline for a successful quantitative study which rather than confirming previous findings will allow truly novel insights. It should be remembered that despite its achievements, quantitative research in this area has employed the same methods and variables for a considerable time. This has left a clear gap in our understanding and measurement of university Knowledge Transfer and calls for a new coherent strategy to ensure novel and interesting outcomes. Hence, the scope and data have to provide additional and supplementary insights rather than repetition of previous approaches. A novel and coherent research strategy is essential to ensure a novel approach, and data sources, methods and potential analysis strategies need careful consideration.

4.1 EMPIRICAL SCOPE AND DATA SELECTION

An appropriate scope must be defined in the context of a quantitative investigation of university-industry Knowledge Transfer. This empirical research focuses on a European technical university and its proximate industry surroundings as a relevant scope for this work. It is relevant for the following reasons: First, the slightly exploratory nature of the rather technical research objectives (RO2a,RO2b) means that thesis scope must be manageable and allow for in-depth investigation. The scope allows in depth assessment of novel methods which is important since the performance of the method and the technical findings are uncertain. Second, the scope of a European technical university allows for interpretation that can be generalized to other technical universities. I chose a Scandinavian technical university with enrolment of more than 11,000 students, a good academic ranking (Times Higher Education Rankings 2017 top 100 , Leiden Ranking 2018[1]) and around 4,000 researchers (including faculty members and PhDs).

These characteristics are fairly average in a European and Scandinavian context, and increase generalizability to other technical universities. The goal is to provide a proof of concept and then, if it proves to have empirical value to make it broadly applicable. Third, a technical university has a high likelihood of the outcome of interest: university-industry Knowledge Transfer. To increase this likelihood, the chosen technical university has a high number of industry-collaborations and commercial activities (Balconi and Laboranti, 2006). This choice ensured that the university research is relevant to industry actors (Leiden Ranking 2018 - collaboration with industry), and thus more likely to be utilized (Schwarz et al., 1998). This allows an understanding of the university structures , and detailed alignment of methods and data sampling.

Given the objective (RO2a) to investigate the potential of a novel measurement, it was crucial to strategically identify appropriate, accessible and consistent potential data sources. To ensure a coherent approach, systematic investigation of potential computational methods was carried out. This was guided by

1 http://www.leidenranking.com/ranking/2018/list

previously developed approaches in empirical work in the field of computer science and computer linguistics (cf. chapter 6). A strategy based on several distinct methods was employed based on the combination of several statistical computational methods rather than the combination of qualitative and quantitative methods.

To ensure novelty, I chose text mining based methods to create a Knowledge Transfer measurement. The decision to work with text mining tools and methods was based on the notion that large amounts of explicit knowledge are captured and achieved in the form of text documents. This is not a new development: knowledge has been passed on in texts since the beginning of writing (Olson, 1996).

Today, the numbers of texts and knowledge sources and their availability are unprecedented (R. Feldman, Sanger et al., 2007). In addition, there are standardized text forms such as scientific papers, patents, technical reports, etc. which are generated specifically to display or transfer knowledge. This made the selection of text data a natural choice. These facts enabled my decision to employ a combination of novel and traditional (text) data sources. To address the thesis research aim I conducted three main strategic steps which are described in the studies in Manuscripts I, II and III.

First, a bibliometric driven study of data structures and research output was conducted to obtain a complete overview of the university research and its internal structures. This was necessary to understand the structure of the university's knowledge system and the changes it had undergone (RO1) (Fung and Wong, 2017; Moed et al., 1985). By focusing only on sources that could provide text data on research knowledge, this study investigates traditional and potential new data sources to obtain a better understanding of data coverage, missing data, data limitations and data potential (Manuscript I). This first step towards identification and an enhanced understanding of the data representing the university knowledge system allowed identification of which data best represent the university knowledge system (for detailed extensive information on data sampling and structures see chapter 7).

Second, Manuscript II is based on the outcomes and insights from the first study. This second study is the core of the thesis, and develops a novel approach to detecting university-industry Knowledge Transfer (RO2a) and evaluating this detection method (RO2b). To identify Knowledge Transfer methods requires the knowledge to be received, detected and verified. As already described, the main data for this study are text data from different data sources. In data and analyses are based on in from Manuscript I and its generated insights, while for the receiver side (companies), a variety of data sources could have been selected. However, not all potential data sources are suitable, accessible or contain sufficient information for the objective of this second study.

It is difficult to find adequate textual representations of company knowledge since in order to maintain competitive advantage companies do not display all of their knowledge publicly. This suggested the need to approach the companies to obtain non-publicly available, more detailed and complete knowledge data, or obtain a less complete but publicly available data. The first option would restrict potential data size substantially. This would present a problem since most computational methods rely on large amounts of data. This meant that the second option of publicly available data was more feasible and we identified company websites and online documents such as annual reports, product descriptions, and company register data as potential sources. For reasons of availability and the amount of data provided on company-websites they were seen as the most appropriate choice (Kayser and Blind, 2017). Can company knowledge be captured on websites? Companies that are research intensive, that use cutting edge technologies need to attract shareholders and convince customers of their leading role in their respective industry. Knowledge and expertise represent competitive advantage which needs to be displayed to attract potential consumers, users and shareholders (Kinne and Axenbeck, 2018).

In many cases the company might use certain knowledge but not display it publicly. However, a large sample is more likely to indicate whether this is a valid approach. To measure or to detect actual Knowledge Transfer, requires examination of the content of text data. This can be achieved using various com-

putational methods. The goal of this study is to identify the best performing methods and tools for this purpose. Third, as a final step, the methods developed and tested in the second study (Manuscript II) need to be validated and improved. Manuscript III identifies potential research output characteristics and parameters that might influence Knowledge Transfer rates on the basis of the novel methods (RO_3), and proposes additional methods to improve the results.

The relevant potential parameters are derived from the findings of the first two studies by extracting common features in their results. The elements considered relevant in terms of data are based on the findings in Manuscript I, and those related to patterns are based on the Knowledge Transfer detection findings in Manuscript II. The key drivers or indicators of the knowledge that has a higher probability of being transferred, may very well vary according to academic field, type of university publication and industry sector. In summary, Manuscript I is mainly a data investigation and focuses on bibliometric data, but is the base for the data sets used in Manuscript II and Manuscript III which overlap and are based on the requirements of the text mining methods. Manuscript II and Manuscript III also apply mostly the same technical methods.

5 MANUSCRIPT I

This paper is based on the published short paper presented at the Science, Technology and Innovation Indicators (STI) conference in Leiden 2018 (see Appendix E) (Woltmann et al., 2018). It is the basis for the later data usage and helps to understand potential shortcoming and gaps in the data structures that are available.

University Research Knowledge Structures and Their Development Over Time

Sabrina L. Woltmann, Melanie Kreye

3rd December 2018

I.1 INTRODUCTION

Universities today are under huge pressure to contribute to society and economic development by supplying novel knowledge generated through their (publicly funded) research (D'Este and Patel, 2007; Tijssen et al., 2009). Universities are seen as knowledge centers, which create new ideas and knowledge (Ankrah et al., 2013; Perkmann et al., 2013), provide novel solutions and drive (industrial) innovation (Cohen et al., 2002). Hence, one of their main functions is strategic knowledge dissemination. Universities achieve this via various activities including: the production of research outputs such as publications and patents; social outreach projects; and direct industry collaboration. Successful dissemination is measured using proxy indicators, some of which focus only on the research output such as (journal) publications produced by academics for other academics (Tijssen et al., 2002; Waltman, 2016), others examine research outcomes related to university-industry and university-society impact (Drucker and Goldstein, 2007). Most empirical studies focus on either the academic or non-academic implications of public research and overlook their inter-relations which potentially underestimates their true contribution (Cohen et al., 2002).

To try to remedy this shortcoming, this study explores the overall systemic structure of university research output based on the example of one European technical university. Universities develop knowledge systems composed of diverse written research outputs (Geuna and Muscio, 2009; Jensen et al., 2003). We aim to examine the development of this internally generated knowledge system. Our investigation spans developments over a decade to identify structures and changes in the university's knowledge system based on its written research outcomes. Our approach identifies: a) the (changing) distribution and composition of the research output system, b) the interrelations among different dimensions of the research outputs, and c) changes in dissemination strategies. It is important to understand what insights may be overlooked by focusing solely on the common (established) perspectives, indicators and databases. We provide in-depth insights into the developments at university to derive

implications for the promotion of university knowledge systems and directions for further empirical research.

For the purposes of the present study, we combine different databases and process novel additional data to generate a coherent and holistic data structure. The analysis employs basic statistical approaches. This methodological framework is used to identify changes over time, taking account of time cumulative data. The empirical results should be interpreted in the context of current empirical research in the field of university Knowledge Transfer and bibliometrics studies. Our findings suggest that a focus on established perspectives and indicator sets is valid if its limitations are taken into account. However, in the context of small case studies, in particular, crucial aspects of university knowledge structures may be missed if other indicators and databases are not exploited. This applies also to studies that focus solely on specific academic research fields which can lead to results with high selective bias.

1.1.1 Background Literature

The composition of university research outcome in terms of written outputs is an important aspect for many academic scholars including, especially, research on public policy, innovation, university-industry collaboration and knowledge management. The literature in these areas investigates the composition of university research structures from various perspectives (Liyanage et al., 2009) which has provided a highly interdisciplinary, but also separate body of literature on university research knowledge (Gherardini and Nucciotti, 2017).

Some studies focus on the economic and/or societal implications of the university research, particularly using econometric concepts, models and indicators (Drucker and Goldstein, 2007). The economic aspect of has traditionally attracted most attention of scholars. The empirical and indicator bases of this research are now well established (Cheah, 2016; J. Thursby and M. Thursby, 2002).

However, the findings vary, but tend to highlight the importance of university research and university collaborations. Work in this stream almost exclusively uses commercially relevant indicat-

ors such as patents and university spin-out activities (Audretsch, 2014; Cohen et al., 2002; Érdi et al., 2013). Some studies focus rather on academic indicators such as patents and/or co-authored publications between companies and universities. The other literature stream (relevant to the present study) focuses on the academic perspective of university research considering the exchange within the academic community (Tartari et al., 2014). The academic community is considered the main recipient/target of academic research (Waltman, 2016). This literature strand uses well developed and established metrics and indicators which include publications outputs, citation counts and sometimes also altmetrics (Erdt et al., 2016). These indicators tend to be based on commonly established bibliometric databases and limited to certain types of research output such as journal papers and book chapters. This contributes to an overall fragmented picture of empirical work on university research impact assessment.

This division into two empirical research streams (econometric and academic) has resulted in a certain detachment of these connected streams, and shows that current indicator frames are rarely being critically assessed or expanded. Some studies investigate both of these aspects together and use overlapping indicators to generate a more holistic picture (Cheah, 2016; Salter et al., 2017). For example, studies that use academic and industry co-publications as indicators where university and industry interests clearly are interlinked. Other developments in this direction include Roach (Roach and Cohen, 2013) and Magerman (Magerman et al., 2015), who combine indicators for patents and publications, generating 'patent-paper pairs' which contain a commercial aspect and relates it to its academic counter part.

These studies achieve a different picture of knowledge flows and knowledge structures (Huang and Zhao, 2009). Most still use related conceptual logics and employ citations and references to indicate prior knowledge and link it either to academia and/or industry. This suggests that there is a need to extend the current perspectives. Our study builds on the above work which demonstrates the importance of the academic and economic impacts of research publications and other research outputs to disseminate knowledge; we expect that the transfer of university knowledge will depend on the structure and composition of the research

knowledge outcome system. We investigate the potential of and gaps in traditional indicator frames. We draw on the two literature streams mentioned above to obtain a detailed picture of the knowledge structures and data availability in one institution. This allows consideration of the overall relevance of university research outcome. We want to understand the internal knowledge system and the importance of certain research output *"dimensions"*. These dimensions refer to the combination of empirical data in the literature on university-industry collaboration, bibliometrics and databases. Our investigation dimensions includes:

1. *Publication Types* of research output, that is, the different text documents or similar that can be considered to be academic outputs. We adopt the classifications in the university's own online database (ORBIT):
    - Journal papers;
    - Conference papers;
    - Book chapters;
    - Newspaper articles;
    - Non-textual; (e.g. videos - may have an additional text description)
    - Net publications;
    - Working-papers;
    - Patents;
    - Memoranda;
    - Other (not otherwise defined);
2. *Research fields* which correspond to university departments/faculties. These include 20 different scientific fields plus one for items with no clear disciplinary association.
3. *Accessibility* which can be Open Access or Closed access (subscription based), the latter referring to articles available by paying a journal subscription fee. In our study, Open Access includes full text available in the university database.
4. *Indexed and non-indexed* entries, or whether the output is included in a publications/ citations database (in our case Scopus).

We distinguish also between traditional and non-traditional publications. In this study, "traditional" includes conventional publications such as journal papers, book chapters, etc., which are indexed in the publication database Scopus and are subscription based (restricted access). Non "traditional" includes publications that are other types or not indexed in Scopus or Open Access publications. This allows us to identify differences between traditional bibliometric data and our additional data.

Our study aims to investigate the structures of a university research knowledge system and their changes, based on academic outputs, using the above mentioned indicators and dimensions. First, we need to understand the structure of the knowledge output system and the overall composition of the knowledge outputs. This requires an analysis of the data structure and its development of the university knowledge system and its variation over time. This is relevant since university research knowledge systems evolve according to strategic and policy changes, which affect the relevance of certain output dimensions. The literature demonstrates the need to address two basic aspects of the knowledge system through available and utilized data: First, the need to understand the general composition of the indicators and related data, and second, changes in their composition over time. This will allow some conclusions on developments in research-based outputs along the different dimensions. We investigate significant changes, meaning increases and decreases in these outputs. We assume that:

Assumption 1 (A1): *Significant changes in terms of the distribution of the dimensions research types, research fields and accessibility occur in research knowledge system over time.*

Investigating A1 will provide insights into the general composition and changes to the outputs of the university. However, it we need also to go beyond ratios and examine actual difference among research output dimensions. This will reveal gaps and blind spots in current empirical approaches through data coverage and quality. We need to determine the meaningfulness and comprehensiveness of traditional research output indicators and data sets, and whether supplementary indicators and data sources are needed. We follow the notion that:

Assumption 2 (A2): *There are gaps in the coverage of traditional research output indicators and data sets which could be filled by the inclusion of supplementary data.*

However, changes to the composition of output distribution patterns and data coverage are not the only relevant parameters. We need also to investigate additional non-traditional data and evaluate their potential. Data coverage is important since pure numbers (see A1) may not show whether the relevant dimensions are represented in an adequate manner. We need to establish the relevance of non-traditional indicators and within the overall system.

Assumption 3 (A3): *Non-traditional research outputs are important and relevant to the composition and structure of the research output system.*

1.2 DATA AND METHODS

Verifying A1 requires several tests to check the different dimensions and to provide thorough results to allow our prediction to be supported or rejected. A clear strategic methodological approach to our assumption is needed to ensure that all relevant aspects are considered. Our data choices, data collection and data processing need to be thorough.

1.2.1 Methods

This study uses common statistical approaches and methods to examine the frequency distribution ratios. We employ χ^2 tests to determine significant differences between the expected and real frequencies in one or several categories. The χ^2 is a commonly used test in bibliometric and econometric studies (Lawani, 1986). We also employ a basic network analytic approach to investigate A3 and to understand the system structure and identify changes over time. Other scholars have used these methods to identify linkages and emerging topics in various scientific areas (Su and Lee, 2010; Zhang et al., 2012; Zhu et al., 2015). We adopt this empirical strategy based on its proven utility to assess research

developments, research quality and knowledge flows in an academic context.

To test A1, we examine overall numbers for the different years (2005-2015) in terms of changes in dimensions 1-3, which include outcome types, research fields and the output accessibility.

For A2, we investigate the three dimensions in their relation to the fourth dimension: we want to identify whether information on the different research types, research fields and access types are equally available in Scopus. This is to see where the traditional data sources might have gaps in coverage. Hence, we are identifying the lost information if a researcher relies only on the most commonly used databases.

Last (for A3), we apply basic network analysis to distinguish the different dimensions in terms of network parameters, such as average degree of the respective types, research field and accessibility. These measures show the centrality and potential importance of each dimension in the internal knowledge network and allow us to see the relevance of in Scopus missing entries.

1.2.2 Sample and Data

We focus only on research outputs collected from one European technical university that could be used for knowledge dissemination. A technical university is a suitable setting for our inquiry since it provides access to basic and applied research and engages in many collaborative research agreements which can result in less conventional outputs (Schwarz et al., 1998). This is then relevant for the university impact (see Section I.1). We recognize that one university is a small research knowledge system and constitutes a very small-scale data sample. However, our study is a first attempt to study the role of and interrelations among different university output dimensions. We exploit quantitative data since our study questions the current notion about university output and data sources. The three main data sources used are: 1) the technical university's internal publications database (ORBIT[1]); 2) the Scopus[2] publications database which contains citations data;

1 http://orbit.dtu.dk/en/
2 https://www.scopus.com/home.uri

and 3) the PATSTAT[3] patent database which provides full-text patent descriptions.

The data sampled from these sources were upgraded through the addition of full-text data and further extensive data processing. The main data set for this study is based on the university's own publications database (ORBIT) which includes all outputs from university employees. Our sample includes all outputs registered in the years 2005-2015 which, according to the database managers, is the period with the most complete and relevant data. We collected comprehensive data on each entry. All entries in ORBIT are linked to meta-data such as a type label, which enables us to distinguish among different output types such as patents, papers, book chapters; a scientific fields label; complete title, author names and abstract for each item.

The final data set includes information on 77,920 entries including some 500 patents (families[4]) assigned to the university. There are 20 scientific research fields in ORBIT including Nanotechnology, Biosystems, Photonics, Space Research, and more traditional fields, like Mechanical Engineering or Civil Engineering. From PATSTAT, we obtained the full-texts of 328 patent applications including references, citations and titles of non-patent literature (NPL[5]). We used these full-texts to identify references to university publications. By combining the data from ORBIT and PATSTAT we retrieved and integrated citations and the other, above described, meta-data. The Open Access publications provided full-texts for around 20,000 publications. We used these to identify references to university publications which we combined with our ORBIT data to identify and integrate further meta-data. This enabled us to generate a very extensive picture of the university outputs and the single items.

Scopus is frequently used for research due to its extensive and reliable content (Boyack, 2015; Kamdem et al., 2017). It provides meta-data and citation data for a great amount of scientific publications. In our case, 28,734 entries overall from the university

3 https://www.epo.org/searching-for-patents/business/patstat.html

4 A patent family is a collection of patent applications covering the same or similar technical content (https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families.html).

5 The rest could not be identified in the common patent database for yet unknown reasons

database could be identified in Scopus. However, from the 23,000 Open Access (full-text) items in ORBIT only around 7,000 were identified in Scopus. To identify university publications, we used title string matching via the Scopus application programming interface (API) since ORBIT does not include identifiers which could be used to match Scopus entries.

I.3 RESULTS

For each our assumptions, we present the results for the relevant dimension. Some tests involved combining two dimensions and we split the findings according to their relevance and interrelation.

I.3.1 Results for A1

To verify our first assumption, we focus on the information collected from the university's internal database. We looked at the distribution and compared frequencies. The most interesting dimensions are research fields and output types. We analyzed the distribution of the different output types in the composition of the university knowledge system. They present a highly skewed distribution with peaks for journal articles and book chapters (see Figure I.1). The difference among types varies from a couple of entries to several thousand and shows clear prioritization of traditional publication types (journal publications, books and conference papers). The ratios do not change significantly in the overall distribution during the period 2005-2015 (see Appendix E).

However, some types show an increase from their own baseline over the years, and change significantly in terms of their composition (see Table I.2). For the research fields dimension, ratios vary widely among fields, with some having only a couple of hundred entries and others adding up to several thousand. This can stem from a variety of reasons: the different numbers of employees in the respective university departments and differences in time spans among different research fields since some were new, and others long standing and other external factors. Some

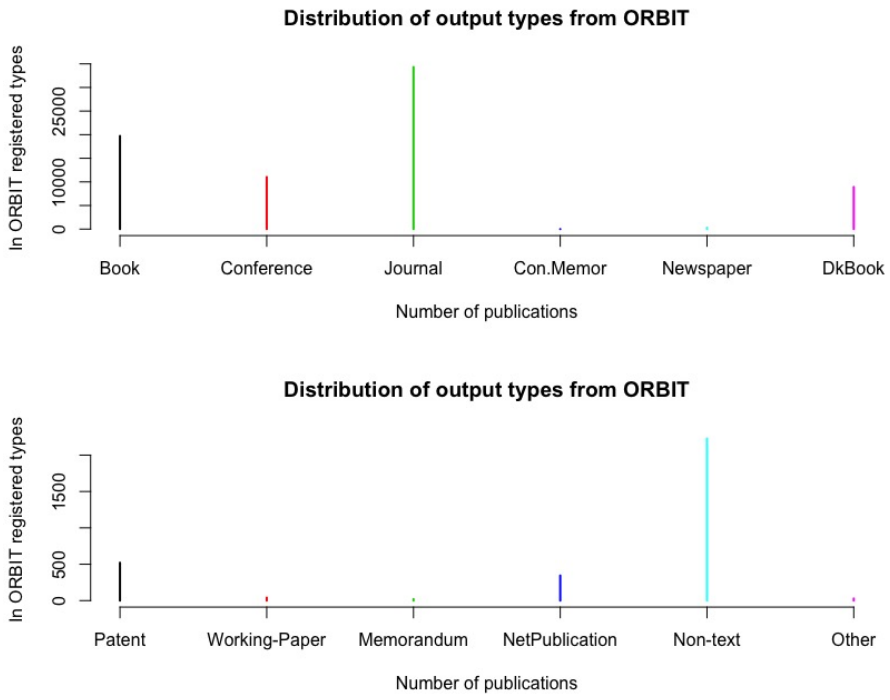


Figure 1.1: Sample of the university publication types

fields show an significant increase or decrease from their baseline over the years and leading to significant changes to the overall composition of research fields in the data.

If we observe the distribution over years in terms of the different research fields, some fields and certain non-traditional output types show increasing relevance, such as *net publications* or *newspaper articles*. We want to discover whether these ratios have a decisive impact on the data and hence on the system structure. To assess how large is the difference, we ran a χ^2 test, which allows us to test the null hypothesis stating that the proportions in several groups are the same, or are equal over the years. Inter group comparison of research fields and output types in ORBIT shows $\chi^2(220, N = 77516) = 20417$ with a p-value $= 2.2 \times 10^{-16}$. The p-value is smaller than the 0.05 significance level we set; hence, we can state that the output type is dependent on the research field.

The final dimension relevant to our first assumption (A1) is the dimension of accessibility. We examined the overall and the yearly distribution of Open Access entries over the years. It is

clear that Open Access items account for almost a third of the university's total research output with an increasing tendency (see Table I.1). The distribution over years is especially interesting and the Open Access changes significantly over time. It became relevant only in 2011 and since then increased up to 2015 to exceed subscription based publications (see Table I.1). Subsequent investigation of the different research fields in terms of accessibility, provides additional insights. We find that this dimension has different importance according to research field.

Publication year	Total publications	Total OA	Ratio
2005	6046	587	0.10
2006	6368	823	0.13
2007	6957	916	0.13
2008	6822	1108	0.16
2009	7183	1533	0.21
2010	7050	1614	0.23
2011	7409	2781	0.38
2012	7491	3305	0.44
2013	7665	3385	0.44
2014	7647	3464	0.45
2015	7789	4197	0.54

Table I.1: Publication availability distribution over years, from (Woltmann et al., 2018)

To test A1, we ran χ^2 tests on the number of publications for each output type and each field (or department/faculty) during 2005 to 2015. We assume that if each output type is distributed in the same proportions during those years, we can expect the distribution of each output type to comply with a ratio scaled by the total number of publications in a given year. For example, if 10% of the publications in the total sample are journal articles, we can expect 1,000 contributions to journals in 2005, if 10,000 publications were recorded for 2005. Applying the χ^2 test between our data and their expected values will determine if the data are independent of the expected values or not. In the case where the data and our expected values are independent, the null hypothesis is supported, which would mean that the p-value of the χ^2 test will show a lower significance level which

we can determine. We chose p-values of <0.01 as the significance threshold[6].

For output types, if the null hypothesis is supported, this means that output types are not steady and vary significantly over time.

For output fields, if the null hypothesis is supported, this means that the research field's output varies over time.

Table I.2 shows the results of the χ^2 test for each output type. It shows that apart from the *Memorandum* types, the distribution varies significantly over the 10 years of our sample.

Table I.3 presents the results of the χ^2 test for each department. They show that with the exception of *Compute/Math*, the distribution of all the fields varies.

Type	**χ^2 test score**	**p-value**	**p<0.01**
Journal Publication	214.03	1.8×10^{-40}	**Yes**
Conference Publication	313.32	2.37×10^{-61}	**Yes**
Book Contribution	108.61	1.01×10^{-18}	**Yes**
Dk Book Contibution	38.16	3.56×10^{-5}	**Yes**
Patent	147.81	1.05×10^{-26}	**Yes**
Non-Textual	225.56	7.31×10^{-43}	**Yes**
Newspaper Article	51.76	1.26×10^{-7}	**Yes**
NetPublication	199.38	2.17×10^{-37}	**Yes**
Memorandum	15.88	0.103	**No**
Memorandum DK	9.8	0.458	**No**
Working Paper	36.85	6.01×10^{-5}	**Yes**
Other	37.77	4.15×10^{-5}	**Yes**

Table I.2: A1 test: χ^2 test results for departments/fields over the years

In summary the distributions and ratios lead to the following statements: First, the overall ratios of research types is extremely skewed with a clear preference for some particular output types such as journal papers, books and conference contributions. Second, the ratio of most research types evolves over the years with the emphasis on traditional types. This supports our first assumption A1.

6 This is necessary due to the strong assumptions about data distributions we make.

Department	χ^2 test score	p-value	significant (<0.01)
MechEng	51.13	1.65×10^{-7}	**Yes**
Vet	78.09	1.19×10^{-12}	**Yes**
BioSys	163.43	6.36×10^{-30}	**Yes**
Food	207.13	5.23×10^{-39}	**Yes**
MAN	131.66	2.14×10^{-23}	**Yes**
CivilEng	55.12	3.00×10^{-8}	**Yes**
ChemBiochem	91.64	2.53×10^{-15}	**Yes**
ComputeMath	13.6	0.19	**No**
Diverse	4767.45	0	**Yes**
Space	76.18	2.80×10^{-12}	**Yes**
Aqua	56.25	1.84×10^{-8}	**Yes**
ElectEng	120.3	4.41×10^{-21}	**Yes**
EngConSto	1477.09	2.23×10^{-311}	**Yes**
PhotoEng	112.65	1.56×10^{-19}	**Yes**
Chemestry	207.44	4.51×10^{-39}	**Yes**
Nuc	567.86	1.35×10^{-115}	**Yes**
EnviEng	66.25	2.33×10^{-10}	**Yes**
Physics	39.49	2.08×10^{-5}	**Yes**
MicroNano	98.47	1.10×10^{-16}	**Yes**
Wind	2658.43	0	**Yes**
Transport	36.93	5.81×10^{-5}	**Yes**

Table I.3: A1 test: χ^2 test results for research fields (departments) over the years

If we look at different research fields we see that some non-traditional types are becoming more important. Generally, the number of registered output entries varies among different research fields and types and change from year to year. This supports the assumption in the first two dimensions. However, the overall picture shows that many differences are over the total time span minor, but some express significant differences, which impacts statistic approaches used on these data.

I.3.2 Testing the Second Assumption (A2)

To test A2 we are interested in how many outputs are missed if we consider only traditional academic output. We identified traditional outputs as the traditional types and the coverage in the publications databases (in our case using Scopus as the reference). Hence, we focused on the dimension of indexing and investigated the other three dimensions with respect to indexing. The overall coverage of the (Scopus) publications database is important, since it is used frequently in bibliometric and academic studies. In our case, Scopus includes 40% of the university's registered entries. However, the distribution of indexed entries shows an interesting trend: as the overall number of entries increases slightly over the years: the number of entries not registered in Scopus decreases slightly, while the entries that are registered in Scopus increases. However, coverage remains at around 40%-45%, which seems low relative to the importance in research of this database (see Figure I.2).

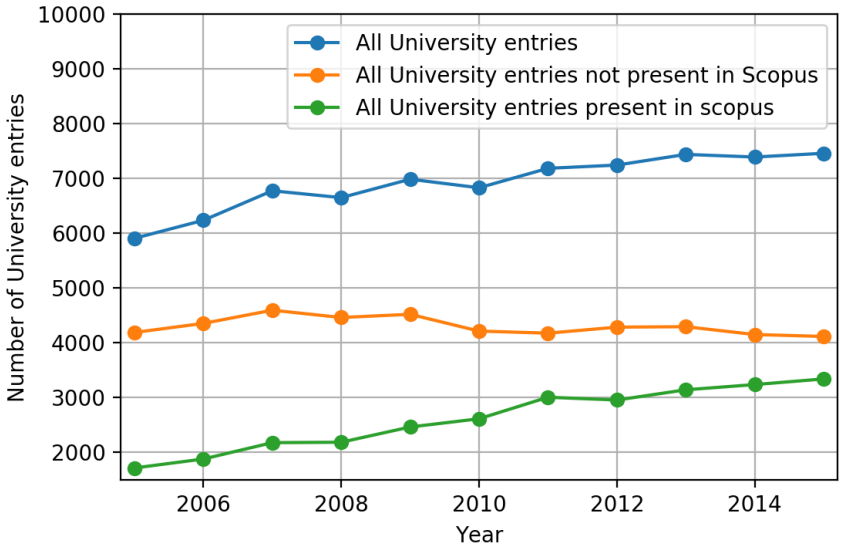


Figure I.2: Sample of the indexed ORBIT entries vs. non indexed entries

Table I.4 presents the results for the χ^2 test on the indexing ratio. The ratio of indexed entries is dynamic and also increases over the years. Here we can see whether the ratios can be com-

pared. For research fields we see that the indexing is again rather skewed towards certain fields more than others, although this depends also on the composition of research types (see Figure I.3).

χ^2 test data			
Year	**Expected**	**Actual**	**Ratio**
2005	2229.79	1717	29.07
2006	2353.98	1881	30.16
2007	2557.45	2179	32.16
2008	2510.26	2187	32.89
2009	2637.1	2465	35.28
2010	2577.83	2615	38.29
2011	2712.59	3008	41.86
2012	2734.86	2957	40.81
2013	2808.85	3144	42.25
2014	2791.11	3239	43.81
2015	2820.17	3342	44.73
χ^2 test results			
Expected ratio			38%
χ^2 score			580.99
χ^2 p-value			2.07×10^{-118}
Significant			**Yes**

Table I.4: H2 test: χ^2 test data and results for indexation over the years

We also looked at trends for the dimension of accessibility, which has become increasingly important in the most recent years (see Section I.3.1). Hence, to investigate A2, we examine Open Access publications only. We want to see whether these items are as well represented in Scopus as the subscription based publications. Figure I.4 depicts this indexing.

It can be seen that Open Access publications are clearly under represented in Scopus, with on average less than a third each year indexed. This is surprising considering the overall increase in Open Access publications. A χ^2 test on the indexing ratio for Open Access papers confirms that the ratios vary significantly (see Table I.5).

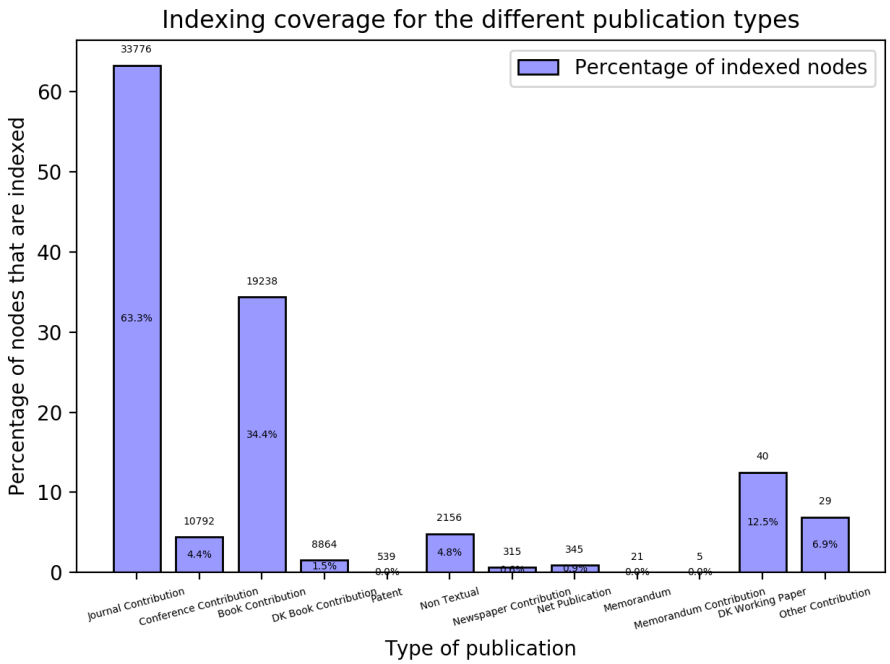


Figure 1.3: Publication type indexing

In the case of Open Access, in particular, the indexation ratio shows wide variation but remains of the same order of magnitude. There is no particular increase observed in the 10-year average of 32%.

We however observe that the indexation ratio increases over the years in particular for subscription based papers although their overall number decreases.

Our results show that Scopus provides diverse coverage of the different dimensions, suggesting that non-indexed data outputs might provide additional insights, which confirms our A2. In particular, the dimensions of accessibility and research field show large gaps in registered data and suggest the inclusion of non-indexed entries in the analysis would capture significant additional information.

Research fields are often investigated individually and it is important to acknowledge the possibility of potential shortcoming in data coverage. Research field coverage in indexed entries suggests that entries that are not indexed might be relevant but are excluded from traditional measures. In particular, the low coverage of books and conference articles is a concern.

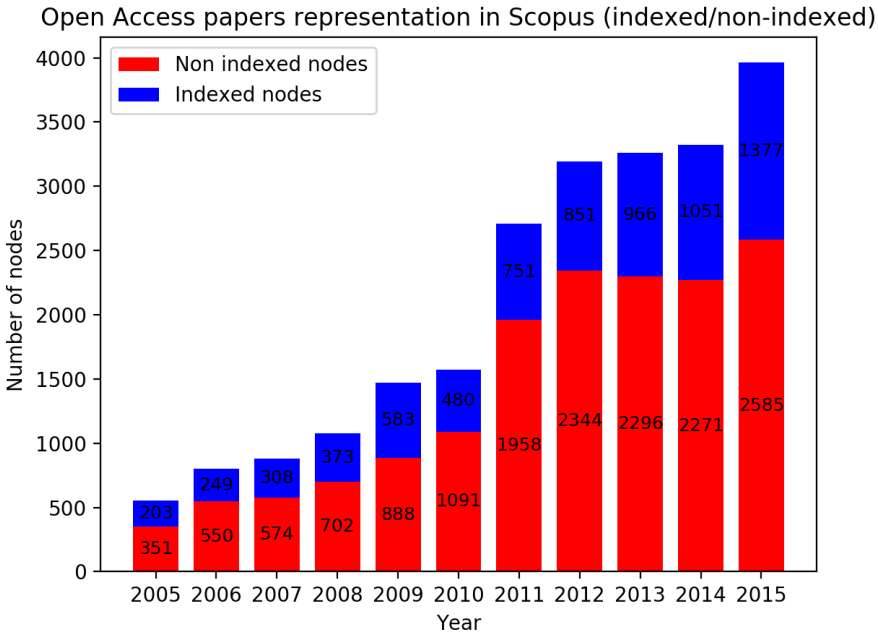


Figure I.4: Open Access indexed Scopus per year

This initial support for A1 and the support found for A2 suggests that a broader empirical approach would be useful, and especially in the case of small scale investigations involving one or a few cases.

I.3.3 Testing and exploration of the A3

To investigate A3, we need to identify the relevance/importance of the non-traditional (in our case non indexed) entries in relation to the overall outputs system. We examine their relevance by using a network approach, which allows conclusions about their importance in relation to overall research output. We build a network to represent the research knowledge system, based on the data collected.

I.3.4 Internal Network & External Network

We started with a common citations network based on the entries in ORBIT that were identified in the Scopus database or that

χ^2 test data						
	Open Access			Non Open Access		
Year	Expected	Actual	Ratio	Expected	Actual	Ratio
2005	174.74	203	36.64	2162.77	1514	28.28
2006	252.01	249	31.16	2196.7	1632	30.02
2007	278.19	308	34.92	2380.94	1871	31.75
2008	339.07	373	34.7	2252.46	1814	32.54
2009	463.97	583	39.63	2228.22	1882	34.13
2010	495.51	480	30.55	2124.38	2135	40.6
2011	854.45	751	27.72	1808.84	2257	50.41
2012	1007.74	851	26.64	1636.32	2106	52
2013	1028.87	966	29.61	1688.44	2178	52.12
2014	1047.8	1051	31.64	1645.2	2188	53.73
2015	1249.66	1377	34.76	1417.74	1965	56
χ^2 test results						
	Open Access			Non Open Access		
	Expected ratio	32%		Expected ratio	40%	
	χ^2 score	95.95		χ^2 score	1366.33	
	χ^2 p-value	3.51×10^{-16}		χ^2 p-value	1.84×10^{-287}	
	Significant	**Yes**		Significant	**Yes**	

Table 1.5: H2 test: χ^2 test for Open Access indexation over the years

could be included through other pre-processing steps. These entries constitute the basis of the internal citations network. The internal network contains only university entries and links (citation) between them. As already mentioned, we can positively identify 28,734 publications in Scopus, but for the remaining 60% of entries that could not be identified, we had to extract links (citations) in a different way. We used the 77,000 titles and tried to link them to the Open Access full-texts. If a title was identified successfully, it was considered a link between the entries. This detection method is computationally intensive and comparatively rigid based on the following constraints:

1. The full title of another publication appears in the second half of the Open Access paper.

2. The title matching is case and punctuation sensitive.

3. The matched title contains at least 50 characters.

These additional citation links derived from the Open Access items help to provide a more complete picture of the output system. We applied the same matching procedure to all registered patent descriptions identifying their NPL. All these entries contribute to the final internal publications network. However, since the expansion is by no means trivial and requires a large amount of additional data processing to ensure a high quality of the results. Accuracy and quality are diminished compared to Scopus, but enable us to integrate in the network additional entries from ORBIT. These additional results should be interpreted with caution. The entries where no citations information or references could be identified were removed from the data for this part of the study.

On the basis of this internal publications network, we also generated an external citations network, based only on additional Scopus indexed entries not assigned to the university. This information is available only for university entries that are indexed. These items were used as supplementary measures of external relevance of the publications combined with overall citations counts of Scopus indexed entries.

Since the third assumption A3 focuses on the relevance of the different outputs, not only the composition or indexing of overall ORBIT data is relevant. Hence we pay attention to the distinction between the different dimensions in terms of their network parameters. We examine measures such as citation linkages (degree of a node) among particular nodes in relation to the rest of the network.

To ensure that this analysis reflects the network in an appropriate manner, we keep only nodes with one or more edge (citation), for both the internal and external networks (i.e. nodes with $\geqslant 1$ in/out edge).

To test our assumption A3 that the added nodes (not indexed in Scopus) are important and relevant, we investigated the changes in the network properties through the addition of non-indexed entries.

The indexed part of the network comprises 28,734 nodes and 49,291 edges, in Scopus between 2005-2015 (1.72 average node degree (citations towards or from an entry)).

The network density is 2.1×10^{-6} which indicates a very sparse network, since density is close to 0. This would be expected in a real network; it improved when removed "lonely nodes" (see Figure I.5).

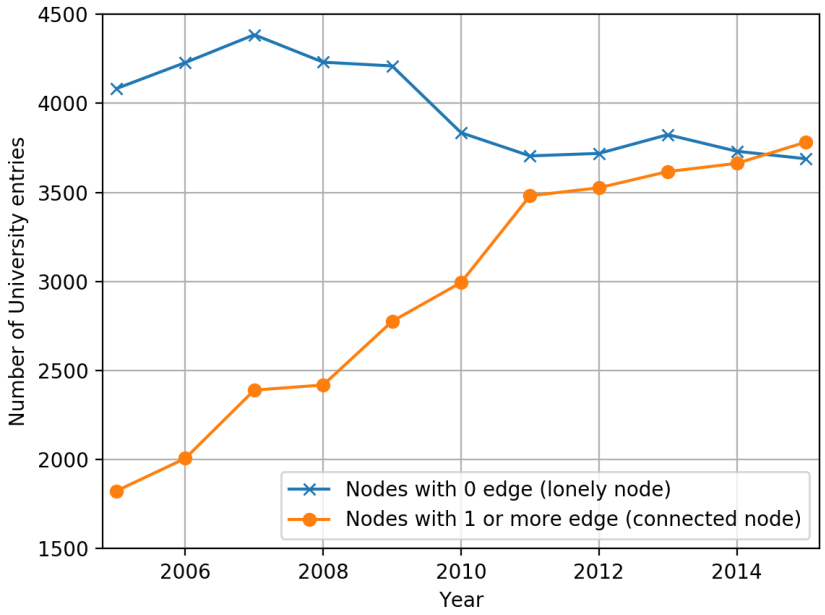


Figure I.5: Lonely nodes for each year

Given the findings from the assumption A1 and A2, it seems that certain research types are relevant for the further process of improving coverage of university entries. But since we could not include all the different types in the internal network (for instance newspaper articles or net publications) we chose one representative type: *patents*.

Our efforts to integrate the patents allowed us to add 43 new edges to the 328 patents, 10 of which were patent-to-patent references and 33 were patent-to-university paper references. These numbers might seem small, but it should be recalled that patents are the most frequent commercial indicators of university innovation. Coverage of this particular and important type was

improved significantly[7]. This approach allowed us to include one of the most decisive commercial indicators.

The other important dimension for the network approach is accessibility.

Given that approximately 25% of university publications in Scopus are Open Access (7,192 out of 28,734), including non-indexed Open Access entries allows additional insights. Also, comparing Open Access to subscription based entries shows that in terms of in-degree citations, they are of comparable importance and quality to subscription based ones (Woltmann et al., 2018).

To check the relevance of non-indexed material, we identified an additional 6,113 edges among 2,657 nodes. These included different output types such as conference papers and other non-indexed publications, such as newspaper articles and online-publications. The improvements to our reference detection for the network are as follows. We know from Figure I.2 that approximately 4,000 nodes per year are not registered by Scopus, and that approximately 4,000 nodes have no edges. The number of lonely nodes decreased to under 4,000 in the five most recent years (see Figure I.5), which is due to our additional reference detection method. The additional detection worked particularly well for more recent papers due in part to the increase in of Open Access papers, where the additional reference detection was possible.

In these kinds of comparatively small networks, such additional information can increase coverage significantly. In Figure I.6, the red line represents data found by the automated detection. The proportion of new edges is large given the conservative setting of the automated reference detection[8].

The Figure I.7 shows the relevance of non-indexed nodes in terms of network properties using the in-degree centrality. It is seen that Open Access publications are performing as well or better than non-Open Access when considering the Indexed Open Access only. The non-index Open Access nodes show more

7 We tried to use the DERWENT collection without success https://clarivate.com/products/derwent-world-patents-index/.

8 This method could probably be significantly improved with some more advanced string detection approaches.

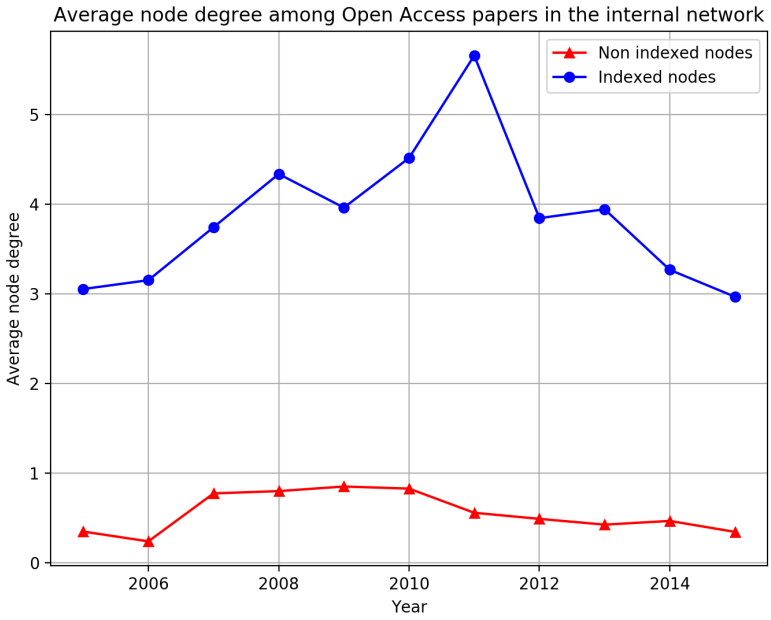


Figure I.6: Average node degree among Open Access publication in the internal network

isolation due to the limited number of edges that were found with the automatic detection.

The final three years in the sample are represented by a dashed line: the in-degrees related to these nodes are less accurate due to relative recent publication dates. Figure I.7 shows that Open Access nodes play an important role in the network. The indexed Open Access nodes are have the same or more citations than the subscription based publications. The non-indexed Open Access nodes also contribute to the network, but are less central. Note that, in general, in-degree centrality is low due to the general network sparsity.

The same results are observed if computing the betweenness centrality of the nodes.

The results for the assumption A3 shows that the non-traditional and non-indexed publications overall have an importance in the knowledge network. This result suggests that for our university case, additional data are certainly relevant to any study. For research types, and in particular patents, as additional data are especially important, since they allow a connection to the uni-

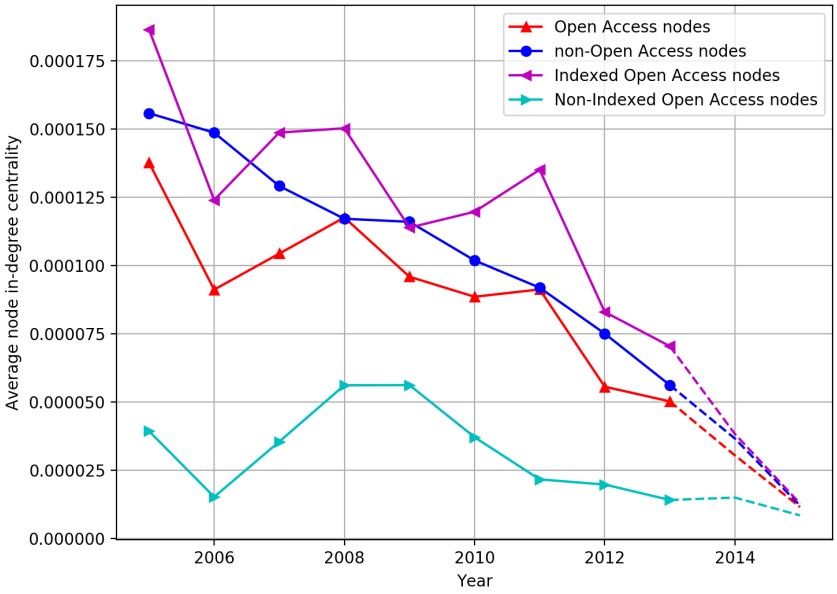


Figure I.7: Average node in degree centrality in the internal network

versity's commercial output and show the interactions between publications and patents. Given the network parameter improvements possible for the dimension of accessibility, it seems that traditional publications miss both numerous and relevant items.

I.4 DISCUSSION

We found overall support for A1, meaning that the research knowledge system undergoes significant changes in its composition over time. It is evident, also, that among research types, traditional types, such as journal articles, conference papers and book chapters, dominate the distribution. The shares of some newer types increased over time, but remain for now at rather low levels. The investigation of research fields shows a high expected yearly variations. However, the findings in terms of Open Access publications shows that their overall share significantly increased in the recent years and ends up constituting more than half of all research items registered. The magnitude of the increase was less expected. Hence, the dimension of accessibility

shows a major shift in the dimensions of the knowledge system structure. This increase clearly reflects contemporary university research publication trends. Our findings show that snapshots of a given point in time of a research knowledge output system can misrepresent the system.

We found partial support for A2, saying that in some aspects of the three dimensions (research type, research field, and accessibility) are underrepresented in traditional (indexed) data sets. This shows that non-traditional data contain some aspects that are left out in the conventional data sources and approaches. The results show that Scopus, as an official publication database, provides a highly diverse coverage for these different dimensions. It is clear that some research types, for instance conference publications, and fields are significantly underrepresented. Again, the most surprising is the dimension of accessibility: coverage of Open Access entries does not show an increasing trend and remains fairly stable in terms of share of indexed items, which shows that indexed entries do not follow internal network composition trends. This leads to changes and uncertainty in the overall coverage.

The results for A3 show that in relation to the structure of the network within the internal university research system, it improved significantly with the inclusion of additional data. In particular the small size and sparsity of the network increased the importance of single items. However, not all relevant dimensions could be tested due to lack of usable additional data sources. Hence, our investigation was limited to the dimension of accessibility, which renders the findings valid, but less holistic. However, in the case of patents in particular, the additional information and their implied connection to the commercial indicators make this feature worth integrating. The amount of edges identified for the dimension of Open Access entries was surprisingly high, which confirms the relevance of this dimension. At the same time, it shows that there is a gap in the system if additional data sources are not included. We also showed that the relevance of Open Access and subscription based items is comparable.

So the dimensions of research types, fields, accessibility and indexation are important in the overall structure of a single

knowledge system. Our findings should be interpreted recalling that they are based on a single case and involved multiple data processing steps, which may be less reliable compared to use of other established data sources. However, the findings give a first indication of structure, changes and developments in a university knowledge system. We can see that certain structural trends seem to dominate and remain more or less stable over time.

1.5 CONCLUSION AND FUTURE RESEARCH

It is evident that a complete overview requires both traditional data and other additional less used data. However, overall, a picture of the actual structure requires more in-depth investigation. The distinction between a general and a more fine-grained exploration is critical for interpretation. However, it is clear, also, that official publication databases include a lot of the relevant data, which is beneficial since these databases are highly reliable and well structured. Large-scale studies can exploit these sources to identify trends. Nevertheless, there are some dimensions that are poorly represented, resulting in some important insights being missed. This will disproportionately favor some investigations by assigning too much weight to some items. We recommend that, wherever possible, empirical investigations should use additional data in addition to traditional standard indicators to close data gaps.

Additional indicators and data can be obtained from novel combinations of existing indicators. Our study should help researchers to understand which dimensions can change and at what level. Hence, depending on the question being investigated, these insights should be taken into account. In the context of small case based studies, crucial aspects of university knowledge structures may be overlooked without the inclusion of additional indicators and databases. This would apply also to studies focused on specific academic research fields and could lead to results with selective bias. However, up-scaling needs to be automated and improved to ensure that in large scale investigations, it is worth the increased effort.

To confirm our findings and ensure that they are not an artifact of our particular data set, therefore it would be useful to study multiple universities' systems. Also, the increasing number of not indexed and Open Access publications suggests the importance of additional data for a complete picture of university output and its impact.

REFERENCES – MANUSCRIPT I

Ankrah, Samuel N, Thomas F Burgess, Paul Grimshaw and Nicky E Shaw (2013). 'Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit'. In: *Technovation* 33.2-3, pp. 50–65.

Audretsch, David B (2014). 'From the entrepreneurial university to the university for the entrepreneurial society'. In: *The Journal of Technology Transfer* 39.3, pp. 313–321.

Boyack, Kevin W (2015). 'Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database.' In: *ISSI*.

Cheah, Sarah (2016). 'Framework for measuring research and innovation impact'. In: *Innovation* 18.2, pp. 212–232.

Cohen, Wesley, Richard R Nelson and John P Walsh (2002). 'Links and impacts: the influence of public research on industrial R&D'. In: *Management science* 48.1, pp. 1–23.

D'Este, Pablo and Parimal Patel (2007). 'University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry?' In: *Research policy* 36.9, pp. 1295–1313.

Drucker, Joshua and Harvey Goldstein (2007). 'Assessing the regional economic development impacts of universities: A review of current approaches'. In: *International regional science review* 30.1, pp. 20–46.

Érdi, Péter, Kinga Makovi, Zoltán Somogyvári, Katherine Strandburg, Jan Tobochnik, Péter Volf and László Zalányi (2013). 'Prediction of emerging technologies based on analysis of the US patent citation network'. In: *Scientometrics* 95.1, pp. 225–242.

Erdt, Mojisola, Aarthy Nagarajan, Sei-Ching Joanna Sin and Yin-Leng Theng (2016). 'Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media'. In: *Scientometrics* 109.2, pp. 1117–1166.

Geuna, Aldo and Alessandro Muscio (2009). 'The governance of university knowledge transfer: A critical review of the literature'. In: *Minerva* 47.1, pp. 93–114.

Gherardini, Alberto and Alberto Nucciotti (2017). 'Yesterday's giants and invisible colleges of today. A study on the 'knowledge transfer' scientific domain'. In: *Scientometrics* 112.1, pp. 255–271.

Huang, Xiao-bin and Chao Zhao (2009). 'Application of Text Mining Technology in Analysis of Net-Mediated Public Sentiment [J]'. In: *Information Science* 1, p. 020.

Jensen, Richard A, Jerry Thursby and Marie Thursby (2003). *The disclosure and licensing of university inventions*. Tech. rep. National Bureau of Economic Research.

Kamdem, Jean P, Kleber R Fidelis, Ricardo G S Nunes, Isaac F Araujo, Olusola O Elekofehinti, Francisco A B da Cunha, Irwin R A de Menezes, Allysson P Pinheiro, Antonia E Duarte and Luiz M Barros (2017). 'Comparative research performance of top universities from the northeastern Brazil on three pharmacological disciplines as seen in scopus database'. In: *Journal of Taibah University Medical Sciences* 12.6, pp. 483–491.

Lawani, Stephen (1986). 'Some bibliometric correlates of quality in scientific research'. In: *Scientometrics* 9.1-2, pp. 13–25.

Liyanage, C., T. Ballal, T. Elhag and Q. Li (2009). 'Knowledge communication and translation - a knowledge transfer model'. In: *Journal of Knowledge Management* 13.3, pp. 118–131. URL: http://www.emeraldinsight.com/doi/abs/10.1108/13673270910962914.

Magerman, Tom, Bart Van Looy and Koenraad Debackere (2015). 'Does involvement in patenting jeopardize one's academic footprint? An analysis of patent-paper pairs in biotechnology'. In: *Research Policy* 44.9, pp. 1702–1713.

Perkmann, Markus, Valentina Tartari, Maureen McKelvey, Erkko Autio, Anders Broström, Pablo D'Este, Riccardo Fini, Aldo Geuna, Rosa Grimaldi, Alan Hughes et al. (2013). 'Academic engagement and commercialisation: A review of the literat-

ure on university–industry relations'. In: *Research policy* 42.2, pp. 423–442.

Roach, Michael and Wesley Cohen (2013). 'Lens or prism? Patent citations as a measure of knowledge flows from public research'. In: *Management Science* 59.2, pp. 504–525.

Salter, Ammon, Rossella Salandra and James Walker (2017). 'Exploring preferences for impact versus publications among UK business and management academics'. In: *Research Policy* 46.10, pp. 1769–1782.

Schwarz, A Winkel, Stephan Schwarz and Robert Tijssen (1998). 'Research and research impact of a technical university—A bibliometric study'. In: *Scientometrics* 41.3, pp. 371–388.

Su, Hsin-Ning and Pei-Chun Lee (2010). 'Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight'. In: *Scientometrics* 85.1, pp. 65–79.

Tartari, Valentina, Markus Perkmann and Ammon Salter (2014). 'In good company: The influence of peers on industry engagement by academic scientists'. In: *Research Policy* 43.7, pp. 1189–1203.

Thursby, Jerry and Marie Thursby (2002). 'Who is selling the ivory tower? Sources of growth in university licensing'. In: *Management science* 48.1, pp. 90–104.

Tijssen, Robert, Thed Van Leeuwen and Erik Van Wijk (2009). 'Benchmarking university-industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities'. In: *Research Evaluation* 18.1, pp. 13–24.

Tijssen, Robert, Martijn Visser and Thed N Van Leeuwen (2002). 'Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?' In: *Scientometrics* 54.3, pp. 381–397.

Waltman, Ludo (2016). 'A review of the literature on citation impact indicators'. In: *Journal of Informetrics* 10.2, pp. 365–391.

Woltmann, Sabrina, Lars Alkærig and Carina Lomberg (2018). 'Open Access' influence on University-Industry Knowledge Transfer'. In: *In PhD Thesis*.

Zhang, Juan, Jun Xie, Wanli Hou, Xiaochen Tu, Jing Xu, Fujian Song, Zhihong Wang and Zuxun Lu (2012). 'Mapping the

knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis'. In: *PloS one* 7.4, e34497.

Zhu, Lin, Xiantao Liu, Sha He, Jun Shi and Ming Pang (2015). 'Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry'. In: *Scientometrics* 105.1, pp. 249–260.

6 METHODS

This chapter describes the methods applied in the two studies in Manuscript II and Manuscript III. It includes technical considerations related to the methods used. The main methods are summarized as well as methods whose implementation was not successful. The primary (mathematical) definitions and considerations of the applied methods are not extensively described since they are included in the relevant studies. However, to ensure understanding, in some cases some aspects of the specifications are included in this chapter. The methods employed for the study described in Manuscript I which focuses on aspects related to the structure of the university knowledge system (RO1), and involves different data, different methodology and data sampling procedures. To avoid confusion and allow for a precise methodological linkage between Manuscript II and Manuscript III, this chapter describes only the methods and data relevant to this, second half of the thesis. However, as mentioned above, Manuscript I facilitates understanding of the data and methodological choices related to the two later studies. The overlap between the methods and data is high which makes their joint explanation appropriate. However, in some cases of specific adjustments, we refer to the studies explicitly in the respective sections of this chapter.

6.1 SUMMARY AND CONSIDERATIONS

6.1.1 Pre-processing

pre-processing is after the data collection the most important step in statistical (text) data analysis. It is key to ensuring good quality outcome from a text mining process, and requires detailed work and adaptation to the needs of the succeeding analyses. The need for pre-processing in text analysis is highlighted by the multiple

problems and the best practice identified (Nogueira et al., 2008). In this thesis, several pre-processing and data cleaning steps were needed to ensure acceptable quality results. In certain cases, some of the steps were repeated to achieve greater improvement. pre-processing was applied to both the academic and the company texts. The main pre-processing steps are based on best practice according to Paukkeri and Honkela (Paukkeri and Honkela, 2010) and Ponweiser (Ponweiser, 2012)[p.33]:

- The retrieved HTML and PDF documents were converted to plain text (unstructured text data);
- The texts were tokenized, meaning word boundaries were defined (as white spaces);
- Conversion of uppercase to lowercase characters (has negative impact on abbreviations/acronyms);
- Stopwords are removed, meaning common words that do not carry content information (in some cases more extensive word removal is used through dictionaries);
- Words were "stemmed" which reduces them to their morphological word stem;
- HTML tags were removed;
- Numbers and special characters were removed (which imposes constraints especially in the context of chemical or mathematical formulas).

Following this the text documents were combined into text collections, text corpora, which formed the basis for the later analyses. The use of the corpora varies according to the declared objectives of specific tasks. The academic texts were classified into corpora based on their respective research field, with one corpus for each academic field. At the same time each company website (assuming it fulfills the given criteria see chapter7.) builds its own corpus, working essentially to produce a text profile of each company. Note that this strategy often results in a higher number of texts per academic field compared to firm websites. The academic corpora contain often a higher number

of documents but not necessarily also more text data. In both cases I obtained a collection of multiple corpora. This separation was necessary to improve the performance of some text mining tools. However, for certain applications such as identification of industry and company topics, we used a website corpus including all website texts (cf Manuscript II). The decision regarding the composition of the collection had to take account of the the size and input requirements of the text mining applications. Some perform document based detailed extraction of keywords; others summarize the differences among the text documents. This requires that the the corpora are built and preprocessed according to the specifications of the applied method.

6.1.2 Document-Term Matrix (DTM)

For the application of the chosen statistical methods, each corpus was converted into Document-Term Matrices (DTMs). This allows a variety of statistical methods to be applied to text data. These matrices are usually highly dimensional and very sparse (Berry and Castellanos, 2007). For more details on the DTMs see Manuscript II. It is the most common text data conversion. It is a vector matrix which describes the frequency of terms occurring in a document collection (Ponweiser2012). The DTM is generated after pre-processing to ensure that only relevant dimensions of words w and documents d are included. The total number of words left after cleaning is the number of columns W, and the number of rows which are the total documents D remaining. In a DTM the element at m, n is the word count (frequency) of the m'th word w in the n'th document d. For a more detailed description (Woltmann and Alkærsig, 2017) (see Appendix B).

$$\text{DocumentTermMatrix}(d, w) = \begin{array}{c} \\ d_m \end{array} \begin{array}{c} \begin{array}{cccc} & & w_n & \\ \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & & \ddots & \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{pmatrix} \end{array} \tag{1}$$

6.2 ANALYTIC STEPS: TEXT MINING

One of the main constraints when working with large amounts of text data is computational feasibility (the trade-off between speed and outcome) of the applications. It is important for results to be generated within an acceptable time frame to be reproducible, replicable and expandable. To save time, and to allow the all text samples collected to be used, basic methods were performed on a sample of publication abstracts (see chapter 7). Among other things, this allowed identification of the most promising and relevant academic fields for more detailed analysis. More advanced and computationally demanding methods were applied to this sub-samples, to refine the findings from the first step (see Manuscript II). This ensured that the approach identifies and excludes research areas and websites that are prone to error (false positives), and sets appropriate thresholds for later applications. This approach greatly increases the accuracy of the methods applied. Different methods were investigated in depth, and only the most promising displayed in the Manuscripts.

6.2.1 Text Similarity and Content Similarity

Identifying similarities within texts is a very challenging text mining function, and recent progress in statistical applications has led to enhanced performance of these methods. It is possible to identify themes or clustering texts into given or generated clusters, and these approaches are used in several research fields across various disciplines (A. Park et al., 2018). However, text and/or term similarity (especially synonyms) remain problematic for computational text analysis Rus, 2014. We can distinguish between two main approaches: word-to-word similarity (Banjade et al., 2015, 2016) and phrase-to-phrase and text or content similarity. Both approaches are covered in main computational similarity measures.. Semantic textual similarity (STS) measures the degree of equivalence in the underlying semantics of paired snippets of text. Such assessments are trivial for humans but constructing algorithms and computational models that mimic human level performance have proven challenging. These measures can be based on probabilistic or algebraic models, and many

studies use a combination of approaches to achieve improved results. However, these practices are often used to detect paraphrasing, and to account for word synonymy which limits their use to identifying only word-to-word, phrase-to-word or phrase-to-phrase similarity (Rus et al., 2013). Several successful methods are based on supervised learning, meaning that many of these methods are trained and tested on labeled data sets. Labeled data sets are data sets where the single text or word has been correctly identified previously, for instance by humans. This provides algorithms with a guiding parameter which it can use to identify common features (so called training). This means that the researchers have testing and training sets which have associated responses which provide the model inputs in relation to the structures to be identified. These structures can be binary or other kinds of categories.

Unlabeled data is used for unsupervised learning to detect common features in the data in the absence of a response. In this case there are no labels giving indications about the right and wrong identification of features (James et al., 2013). This approach is foremost used for data that is not yet very well understood to identify structures. Supervised learning is generally considered to be superior, and to perform significantly better than unsupervised learning (Hübner et al., 2017). It has been shown to perform well in several areas of text mining and is the most frequent form of text processing in real life contexts. In the case for instance of sentiment analysis, it is used to identify positive or negative opinion content in texts (e.g. user evaluations, comment sections or social media entries) or classification assignment (Xu et al., 2010). Supervised learning has a higher potential in the case especially of particular when one expects certain outcomes which can be categorized as true or false.

However, in the present thesis a supervised approach was not possible due to the extent, length and complexity of the data structures (see chapter 7). This imposed some limitations on the evaluation of the methods applied since it was not possible to some standard parameters such as rate of false negatives and the number of true positives in the total data set that could not be identified. Hence, performance was measured only by the results of the human verification which in turn, involved signi-

ficantly less data work. An additional possibility for many text mining applications is natural language processing (NLP) (Bird et al., 2009) whose use in recent years has produced state-of-the-art results and uses often so called deep learning techniques. Deep learning is based on architectures such as neural networks which are based on dense vector representations. It can be supervised, semi-supervised or unsupervised, and has led to major improvements in many areas of text analysis (Young et al., 2017). However, most of these algorithms need larger and more coherent data sets than this thesis provides resulting in them being discarded in this case. However some minor attempts to exploit them are documented in the following sections. Additionally, the data limitation referred to above made it necessary to try many different methods, most of which had to be discarded due to their limited performance. In summary, for the purposes of this thesis, many established text similarity applications were too narrowly focused on very short pieces of text (phrases or words) and relied mainly on the linguistic composition. I exploited some basic algorithms but combined them in different ways to identify overlapping content for entire documents, while relying heavily on co-word occurrence and other measures.

6.2.2 Term Frequency-Inverse Document Frequency

To examine and compare the two text types (company-websites and publications), I first performed very simple keyword extraction and comparison. Given two documents, one from a website and one from a publication that share the same words, there is a likelihood that they will have the same content. However, simple co-word occurrence did not seem sufficient for such cases because the language and content on company-websites tends to be different from that in academic publications. It was assumed therefore, that an additional treatment would more likely yield usable results. A commonly used method is term frequency-inverse document frequency (TFIDF) indexing which is among the simplest and most frequent algebraic methods in the field of applied text mining (Robertson2004). Although being one of the older methods, it often outperforms some of the newer more advanced techniques. It can be used either as a stand-alone

(indexing) method, or as a weighting scheme for other methods. The basic principle of the TFIDF identifies the most relevant words per document[1] , and it can be used to reduce the dimensionality of a document to a small set of terms aimed at capturing its main content. The basic assumption is that more frequently occurring words in a document are more relevant to its content; however, if these words appear in many different documents, they are less essential to the content of a single document.

This method is applied at the document level and yields keywords at the same level. This is crucial, since many methods are applied to the corpus level which makes their interpretation more difficult. TFIDF generally is calculated by multiplying the word frequency (of the document) and the inverse document frequency, meaning the logarithm [2]of the total number of documents divided by the number of documents that contain the word Manuscript II (Gabrilovich and Markovitch, 2007). The disadvantage of this method is that term frequency might represent only the content of a specific part of the document's text, or might overlook words that might be common in the corpus but nevertheless decisive for the content (Xia and Chai, 2011). Additionally, synonyms or word similarity such as "car" and "vehicle" are not detected using this method, while the context is lost since the main content of the document or the entire corpus is not considered.

However, as a first step TFIDF can give a computationally efficient indication of the similarity between documents. TFIDF generates comprehensive keyword lists for each document [3]. To compare documents I limited the keyword lists to a maximum of 50 words and compared the lists with a common similarity measure: the Jaccard coefficient (for more details the Section 6.3. One to one comparison between academic and company text keyword lists was performed on the keyword lists to identify those with common content. The relevant matches were reviewed

1 Derived from occurrences, meaning that this is not always fully aligned to a human reader's assessment.

2 This is done for comparability in terms of normalization; it is not obligatory but is very common.

3 Recall that whether a method gives an outcome per document or at the corpus level is important.

and verified manually (for more detailed information see Section Error! Reference source not found.).

6.2.3 Latent Dirichlet Allocation (LDA)

To achieve the thesis objectives (RO2a and 2b) I needed to improve the tools used to detect Knowledge Transfer by identifying text similarity using other means than TFIDF because of its limitations described above. Among commonly used text mining strategies, I examined topic modeling (Richardson et al., 2014). Topic modeling is used in various text mining application contexts to derive 'topics', or the different content streams in a corpus Topic models are generally unsupervised algorithms [4]. Topic modeling assigns words and documents into specific (unobservable) topics in a probability distribution (Blei et al., 2003).

This means that every corpus has a certain number of assigned topics which contain words which at corpus level. These can be considered keywords. Every document includes a mix of topics, every topic is comprised of a mix of words, and every document includes an assigned number of topics (Ponweiser, 2012). This helps to separate corpora in terms of their overall topics, and identify their general areas. This is a major advantage especially for websites which are in this thesis case not classified into industries, or scientific areas. But it also provides a more in-depth understanding about academic corpora. This counter-acts the shortcomings of the TFIDF. Topic models can detect content relatedness at a broader level. It allowed me to compare corpora and documents simultaneously which added another dimension to the comparison. The topic models are used also to estimate the proximity of industry and university related themes.

Before selecting a method, I explored various topic modeling algorithms and their properties. Among others, I considered latent dirichlet allocation (LDA), hierarchical dirichlet process (HDP), correlated topic models (CTM) and dynamic topic models (D-TM). The latter two are extensions of LDA and have some

4 I am aware that there are topic models that can be used for classification with semi-supervised approaches, however, they are less frequent and show mainly weak performance https://github.com/vi3k6i5/guidedlda

advantages related to their assumptions. For instance, CTM and D-TM use multinominial distribution instead of dirichlet on which LDA is based, to represent the word space. This means that CTM allows for correlation among topics while LDA assumes independence. In academic texts in particular independence of topics seems extremely unlikely. However, topics are still generated in an unsupervised manner which means that quality is not an objective measure but rather is related to the problem to be solved. HDP on the other hand, assumes the same distribution as LDA but does not require the parameter of topic numbers which LDA needs. However, all the models build on LDA and hence are computationally more expensive, meaning their application takes much longer. This is a major draw back especially considering the needed to adjust parameters and continuously to verify performance. Given the amount of text and the need to adjust thresholds and parameters, computational efficiency is a major factor in the present thesis. Nevertheless, I checked the performance of D-TM, CTM and LDA on a small sub-sample to investigate potential differences[5]. I found differences in topics but few qualitative improvements which led to my choice of LDA based on its advantages. Its application required some decisions about parameters:

1. The number of topics K required to be identified;
2. The changes in the word distribution had to be determined (α);
3. Granularity of the topics needed to be set (β).

According to the literature K can be defined in several ways. One of the most common approaches involves the researcher setting a given K based on his or her knowledge of the data. However, since this project contains several different corpora this strategy seemed unfeasible. Hence, K was defined in two ways. One relates to the academic corpora where the optimal number is calculated using the harmonic mean. This method produces several models with different numbers of K and compares the harmonic

5 In the text mining community it is common to verify the outcomes of unsupervised algorithms and assess their output manually.

means of the log likelihoods[6] . It was applied to the academic corpora to assess the optimal number of topics in a statistical manner. However, this method provides only an approximation, and can only be validated based on the researcher's assessment of whether the outcome is suited to the task at hand. I applied also the cross-validation process proposed by Grün and Hornik (Grün and Hornik, 2011), and this generated mixed and not clear results. Both methods are fairly computationally expensive and hence, not suited to application to a large number of company-website corpora. I decided on a less flexible identification of K for the firms, and set the number according to the document sizes of each corpus (for more details see Manuscript II):

$$\mathrm{Document}_m \geqslant 3000 : K = 200$$

$$\mathrm{Document}_m \geqslant 2000 : K = 150$$

$$\mathrm{Document}_m \geqslant 1000 : K = 100$$

$$\mathrm{Document}_m \leqslant 1000 : K = 50$$

Manual investigation suggested that this was a suitable approach. The hyper-parameters for the LDA in our case are aligned to the need to identify common content rather than to classify a document into a topic. Hence, I used parameters that would give high probabilities to a smaller number of topics per document. I found also that the standard $\alpha = 50/k$ performed well in this context. For more information on determining the posterior probability of the latent variable see B. Grün and K. Hornik (2011). In-depth consideration of the model and the parameter adjustments are included in Manuscript II.

6.2.4 Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA) is a common statistical measure in text mining but has diminished in popularity recently. This method was applied to improve findings of Manuscript II in Manuscript III, but did not perform as well as expected (see section III. Technical and mathematical details are included in

6 In this case I used a function designed by Tyler Trinker specifically for the R environment. See https://github.com/trinker/topicmodels_learning.

the following Section after Manuscript II. It is computationally feasible and has performed well over time in terms of dimensionality reduction and detection of synonyms. I used it as a second potential addition to TFIDF to increase the likeliness of similarity detection. It has the advantage that it reduces the dimension of the entire corpus to a more manageable size, and considers the proximity of words while bases the calculations at the document level.

6.3 SIMILARITY: JACCARD VS. COSINE

Crucial to text similarity are the distance or similarity measures that should be used. There are several measures, and most have particular advantages and disadvantages. In this thesis, they are used to identify document pairs (matches) that are considered "similar," and those that are considered not similar. This is an important process since after extracting keywords, topics, and semantic spaces a comparison or similarity measure is required to find potential documents with matching content. The primary need was to identify a simple measure that would provide an adequate estimate of distance. In investigating the options I discovered that my choice of applications and strategies was limiting the potential similarity or distance measures. I decided to run many of the applications on separate corpora or DTM which resulted in scores for words that were not comparable across applications although they were normalized in most cases. Therefore, I had to consider two main aspects related to the similarity measures: First, I would not always have comparable values for my vectors since they are generated using different matrices. Second, there would be a high level of variance in the length of the vectors. These constraints led to the conclusion that the most frequent measure, cosine similarity, would not be sufficient since it is based on scores. I needed a way to compare sets that took account of overlapping words which suggested use of *Jaccard distance* or *Jaccard similarity* measures (Gomaa and Fahmy, 2013). Jaccard similarity is based on the intersection between two sets. The Jaccard similarity value is between 0 and 1, where 1 indicates the most similar (in this case identical) sets

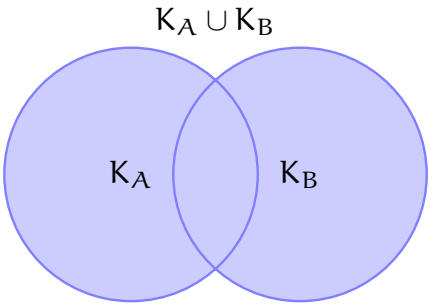


Figure 4: Union

and o indicates no common features (Woltmann and Alkærsig, 2018).

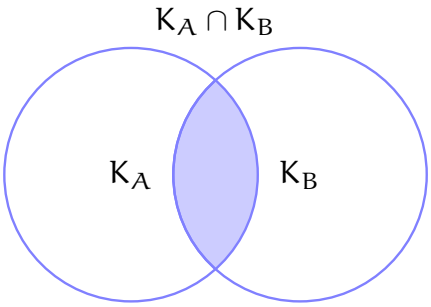


Figure 3: Intersection

Given the set of keywords from one academic document (K_A) and a second set of keywords from one website text (K_B), the Jaccard similarity denoted $J(K_A, K_B)$ is obtained with:

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A|+|K_B|-|K_A \cap K_B|}$$

This measure is most used in Manuscript II and Manuscript III. However, the complexity of the task demanded some minimum thresholds for Jaccard similarity which had to be adjusted based on the findings. These considerations and adjustments are described in C.

For the LSA comparison I chose to compute *cosine similarity* since the values for the vectors were computed to allow comparison, and in certain cases pure set comparison would have

assigned overly high scores to foreign language fragments. Cosine similarity measures the similarity between two vectors which are not zero in an inner product space which measures the cosine of the angle between them. This is a commonly used measure and is adequate for applications of LSA. However, again flexible thresholds need to be set to allow further comparisons.

In the case of both similarity measures the thresholds, i.e. the minimum scores for a document pair to be considered a positive match, had to be adjusted according to the findings. They can depend on various parameters such as academic field, average of the comparison scores or maximum scores. The length of vectors etc. also had to be considered.

6.4 HUMAN VERIFICATION

Human verification is fundamental to Manuscript II and Manuscript III. It is crucial to the reasoning underlying the verification and the thesis research objectives (RO2a and 2a). The overall thesis aim is identifying university to industry Knowledge Transfer with computational methods. As already mentioned, the aim is to detect both overlapping areas of interest and concrete instances of university research being exploited or displayed by a company. To verify that the methods used detect this level of granularity required an additional assessment process. For example, the text mining methods identify some common features, especially the unsupervised algorithms but whether this represents real instances of identical knowledge requires judgment by a supplementary measure. This part of the thesis is particularly challenging since academic texts can be highly complex, and judging the relationship between two texts is difficult even for humans. During the course of this project, several different human verification approaches were tested, and were found to result in different assessment quality and granularity. In one case, verification was assessed by several individuals to ensure inter-category reliability and greater coherence in assessments of

positive and false document pair matches. Three higher educated verifiers were asked to categorize the matches into five different categories C:

- **First category**: Web texts related to a university publication;
- **Second category**: Web texts most likely related but missing an explicit link to a university publication;
- **Third category**: Web texts referring to the same area but a different sub-field of that area;
- **Fourth category**: Text pairs including similar topics but with no deeper connection;
- **Fifth category**: text pairs with no overlaps.

Inter-class coherency was very low in that case Manuscript II, and revealed problems related to this verification approach (see Appendix C). This was a major concern since verification is the final step in the thesis to ensure performance of the methodological approach. After examining the individuals' categorizations and feedback, it was realized that the categories were too fine grained which was causing differences in how they were being understood. To solve this, the categories were regrouped into four new main categories which provided more robust and coherent results and allowed a first reliable assessment of the match building performance. The new categories are defined in Manuscript II as follows:

- **First category**: Identical topic: University contribution
- **Second category**: Identical topic: Potential university contribution
- **Third category**: Common topic: Unlikely directly related
- **Fourth category**: Different topic: No match in content

- **Unclear**: Could not be classified

However, inter-category reliability remained comparatively low although improved. To ensure higher quality and a better final decision about the single document pair category an additional qualitative statement from the individuals was included which referred to their stated level of certainty about the class and content relationship. This approach showed that the assessment was most coherent if the individual declared a high level of certainty related to his or her classification (cf. Manuscript II). This led to a final adjustment for the last study which relied only on the opinions of individuals that declared a high level of confidence about their ratings (cf. Manuscript III). The final Manuscript is based on only three main categories which are relevant to the assessment:

- **First category**: The potential matches, meaning the number of potential document pairs the methods reveal;

- **Second category**: The confirmed matches, meaning every pair that is seen as having identical content;

- **Third category**: The matches with clear content overlap but no direct connection.

All the potential matches that did not fall into the second or third category were considered false positives. This allows a clear distinction which reflects performance and makes the assessment feasible and less complex. Given the complexity of the task this project needed to identify the most feasible and reliable strategy to deal with the verification of statistical output. It is of utmost importance to verify the findings to ensure that statistical outcomes represent the reality. This resulted in the several important strategical adjustments described above.

7 DATA & SAMPLES

This chapter provides a detailed description of the development of the data sets, text samples and databases underpinning Manuscript II and Manuscript III. As in the Computational Methods (cf. Chapter 6). Hence, it considers the second two studies; the first has a different scope and uses different data sets (see Manuscript I). The data, described in this section constitute the foundation to the empirical analysis in the second main part of the thesis. The data and data collection choices are aligned to the computational methods requirements (see chapter Error! Reference source not found.). Data collection is described in detail to ensure understanding and reproducibility. The two studies are based on particular samples, and the studies describe the samples but for space reasons do not include technical details. However, such detail contributes to the understanding of data generation and quality. To facilitate an intuitive understanding of the samples the following section refers to different samples and the relevant manuscripts. Some of the sample specifications are described in full in the manuscripts and hence are not included here .

7.1 DATA AND CONSIDERATIONS

Manuscript II and Manuscript III rely on the same two main data sets which are employed differently and adapted to the respective needs of these studies. The first set of data on university research include academic publications and their meta-data from the university.

Academic publications are commonly used as outcomes and academic-to-academic transfer means, and hence are suited to the purposes of the present thesis (for more details see Chapter 3). The second main data set includes relevant company websites which are representative of the knowledge companies display to their customers and stakeholders. The content of company websites may not display full company knowledge for several reasons: companies may adopt a secrecy competition strategy in relation to novel knowledge and inventions but also may display on such information as they consider relevant to their audience. Hence, this is perhaps an imperfect but a useful starting point to obtaining evidence on knowledge transfer. For both text samples, full-texts and/ or parts of texts were collected and pre-processed.

7.2 UNIVERSITY DATA – PUBLICATIONS

This thesis uses publication data from the university publication database 'ORBIT[1]' which records all publications authored by a university employee. Since ORBIT records all codified output from universities it provides a more complete picture of university publications than large academic databases such as Scopus or the Web of Science (WoS).

It avoids well-known issues related to identifying university affiliation which both Scopus and WoS need to verify authorship by a given university. Registration of publications in internal university databases became mandatory in 2012, and all decentralized publications databases held by individual faculties were integrated in 2013. This might have an impact on the quality of pre 2012 data. All publications recorded between years 2005 and 2015 were extracted, including relevant meta-data such as publication year, author's department/faculty affiliation , full title, a unique identifier (uuid), author name(s) and publication type, journal name and whether the full-text is freely accessible as a (green) open access publication (for more details Manuscript I).

1 http://orbit.dtu.dk/en/

The academic texts used for the text analysis in this thesis are the abstracts, which are directly available in ORBIT. Use of abstracts, as a representation of content is common in the bibliomentric community and has proven to provide good results (Berry and Castellanos, 2004; Patra et al., 2003).

To ensure a comparatively complete picture we removed abstracts containing less than five terms or consisting mainly of HTML tags . However, it should be noted that academic abstracts tend to be short and contain in average less than 250 words (not all of which are content relevant). Therefore, in a few cases additional full-text publications were downloaded to investigate potential performance improvements. Only English language texts were extracted which excluded for instance mandatory Danish summaries of PhD theses, which are also registered as publications in ORBIT. Manuscript II and Manuscript III provide a detailed overview of total numbers, including the (abstract) texts and terms from ORBIT. They provide detailed descriptions of the distribution of research areas in each year (see Manuscript I).Manuscript II and Manuscript III provide a more detailed overview of publication dimensions, including their types, their accessibility format (e.g. Open Access) and their distribution over research fields.

7.2.1 Meta-data

Author's departmental affiliation was used to indicate research area; no other information such as keyword extraction was used to validate this information with the result that highly interdisciplinary research might not always be assigned accurately. Information on department affiliation is highly detailed (and often at the research group level) and contained more than 204 different labels. These labels reflect changes to the structure of the university over time which means that over time some labels disappear and some new ones are added. To use these labels to represent the research field of the publication, we manually grouped them into overall research areas. This resulted in the

assignment of 20 different research fields, and one collection of 'diverse' content that could not sensibly be assigned to any of the 20 areas identified (for the exhaustive list see Error! Reference source not found.The manual grouping was based on information on the university structure (provided by the relevant university administration). This labelling was crucial for the empirical work since differences in research fields and their impact on industry have been discussed in many studies and should not be overlooked. A total of 39 publications authored by administrative staff was removed from the entire sample.

Publication Age (year) is another important variable for the analysis in Manuscript II and Manuscript III. It provides a parameter which might explain better or less good transfer probability for a publication, either because older publications might be less likely to be used or because they have had more time to become established knowledge in the industry. Hence, publication age might play a role.

Whether a publication is an *Open Access* publication is important mainly for the context of Manuscript I and Manuscript III. It is stored as a binary variable (full-text available: "TRUE"/ "FALSE") with the result that for a full-text analysis the document had to be additionally extracted from the database. This parameter is important also since university policies and academic notion change; this parameter changed the most over time. This is particularly relevant to some of the statistical assessments. However, since this is not a commercial database like Web of Science (WoS) or Scopus a small number of tags were wrongly assigned, and in a very small number of cases the full-text attached to these publications contained only a few words.

Title is an important variable and was used to identify research items and citations (Manuscript I) and to identify when full texts were need to assist the manual process (Manuscript II and Manuscript III) However, in some cases titles were too short or too generic to be useful for either of these purposes. In particular, string matching in a context of short character strings is limited

since it makes valid detection nearly impossible (see Manuscript I).

Author name(s) were not used since disambiguation even for an individual university is well-known to be challenging and error prone. However, it could have constituted an additional step in some instances.

Overall numbers- Academic Publications

The total (cleaned) number of ORBIT entries was around 77,000 for the period 2005 to 2015, of which more than 23,000 were freely accessible from the ORBIT database (for more details see Manuscript I and Manuscript III). Overall this university's publication numbers increased only slightly over 2005-2015 (see 5). This was not a surprising finding, but since universities generally aim to increase their publishing rates a higher increase might have been expected.

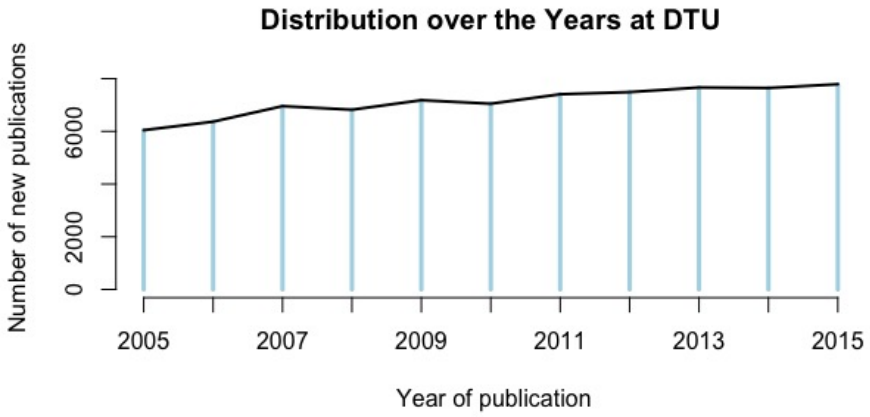


Figure 5: Overall publication rate of the university between 2005 and 2015

However, different departments and research areas show highly diverse overall trends which is surprising considering the generally uniform distribution of university departments. Some

of this diversity might be explained by the emergence of new university departments or enlargements made to existing departments which could lead to a sudden increase in publications in a particular area (e.g. Wind Energy department. Also the unregulated reporting pre-2012 might be having an effect; however, this could not be verified. Another important aspect of department level data is that the total number of publications varies hugely from a few hundred per year (e.g. Nuclear Research) to several thousand (e.g. Compute Mathematics). This matters for comparisons that take account of research area. The other fluctuations might be due to changes to the internal structure of departments and groups such as researcher recruitment or quitting, budget changes or increased/decreased consulting or teaching hours.

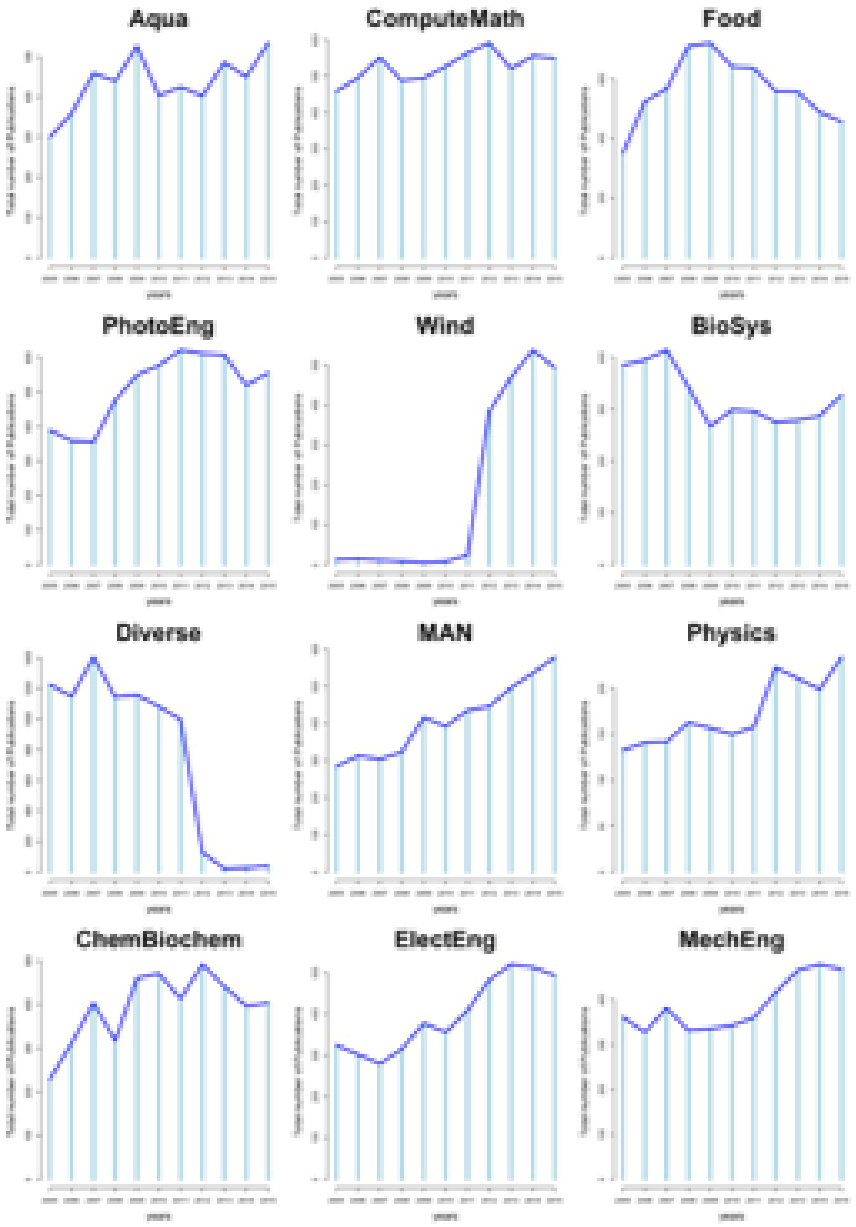


Figure 6: Publication rates for the different university departments between 2005 and 2015

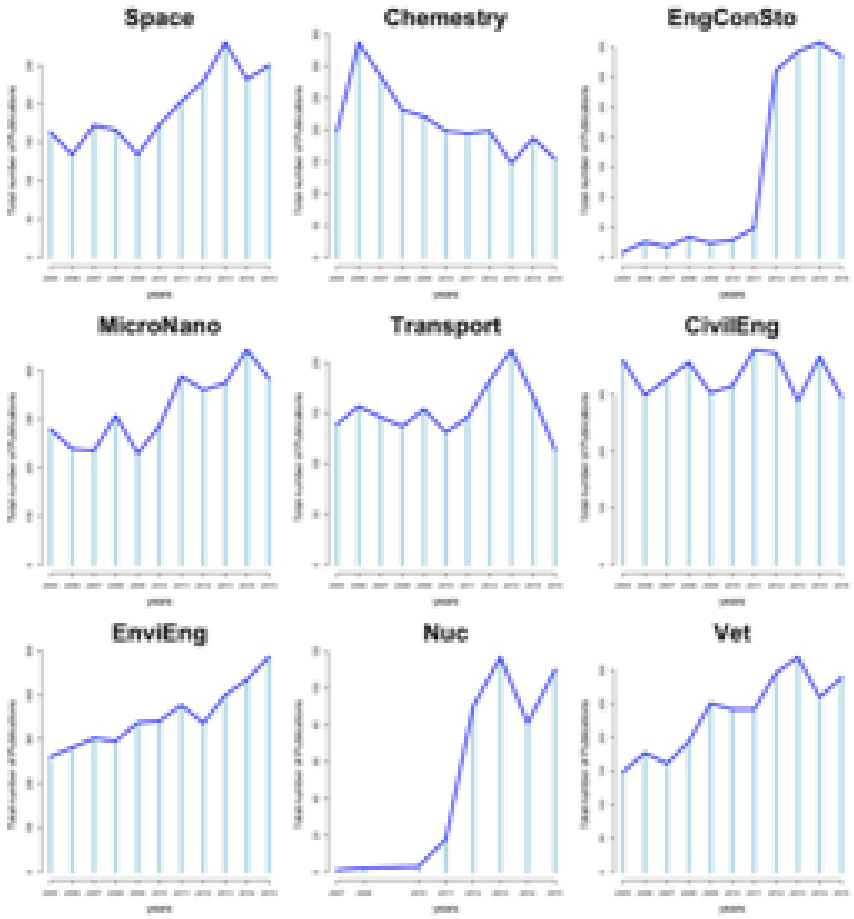


Figure 7: Publication rates for the different university departments between 2005 and 2015

7.2.2 Websites texts

Using information on websites required some additional parameters and criteria to make this approach appropriate for the purposes of this thesis. Use of website information for research is not uncommon and has generated various new insights (Miner et al., 2012). First, due to some basic requirements and performance issues related to text mining methods (cf Chapter 6) a cross lingual approach is often not beneficial which requires that all the utilized websites should be in the English language. Second, the websites must provide sufficient data (terms and pages) to

make comparison feasible. I set a limit of a minimum of five pages in English per website. Language verification was based on using an English dictionary; if more than 60% of the words in the document were in the dictionary the site was considered to be English.

Additional to these main basic requirements the attempt to implement a new measure, a proof of concept, requires some supplementary considerations and parameters. First, the sample should include texts that were likely to represent the transfer of knowledge from the university. This was ensured by imposing the criterion for relevant companies of a 'formal' connection to the university in the form of a contract, or a clear affiliation via hyperlinks on the university website. To ensure a certain level of relationship large online service providers such as Google, Facebook, etc. were excluded from this list.

To include only companies registered in Denmark different identification methods were used (for more details see Manuscript II). Formal partners of the university were collected from an internal database which contains the names of all entities with a contract with a university; this includes companies, universities and institutes during the period 2006-2016. This time period was chosen since according to staff members, the database was more complete from 2006 onward, and university outcomes are available only from 2005 which makes a time lag of one year a suitable transfer time. A few companies had gone out of business, had merged with another company or had generic names which did not allow their identification. The database includes various contract types from research agreements such as industry PhDs, to consulting agreements to pure service provision for instance cyber-security; however, contract type is not considered. Only contracts established between a named researcher or a university research unit were included in order to exclude pure service providers such as cleaning agencies and so on. However, it is possible that separate agreements between individual researchers and companies exist and are not recorded centrally. Identification of companies and assignment of company websites was done manually via an extensive online search.

In a second case (related to Manuscript III) a sample of second degree partners, i.e. partners of formal partners was generated. Second, only companies with a relative regional connection were considered.

Sampling of the websites was performed via web crawling. Using a self-developed tool built in the programming language Python ensured the flexibility needed for data collection . This approach allowed the extraction of a website's entire HTML (Hypertext Markup Language) content based on one base-URL (Uniform Resource Locator), and opening and downloading all related sub-pages for the given base-URL. Since the aim of this thesis required as much company relevant data as possibly publicly available content in different formats (e.g. PDF) was also downloaded and converted to plain text. To increase the efficiency of web-based data collection certain limits were imposed:

- A maximum of 1000 sub-pages per website
- 30 MB as the maximum file size
- A maximum of depth level 2, one from the original base URL (for second degree partners)

These limitations were important to avoid extraction of huge websites such as news websites, or large company websites which included content in several different languages which made them useless for our investigation. Some had several thousand sub-pages.

Overall numbers- Websites

From the contract database and the hyperlinks from the university website a total of 1402 potential entities was identified for the years 2006 top 2016. These represent formal first degree partners. This number does not include other collaborating universities or international public research institutes. However, some of the

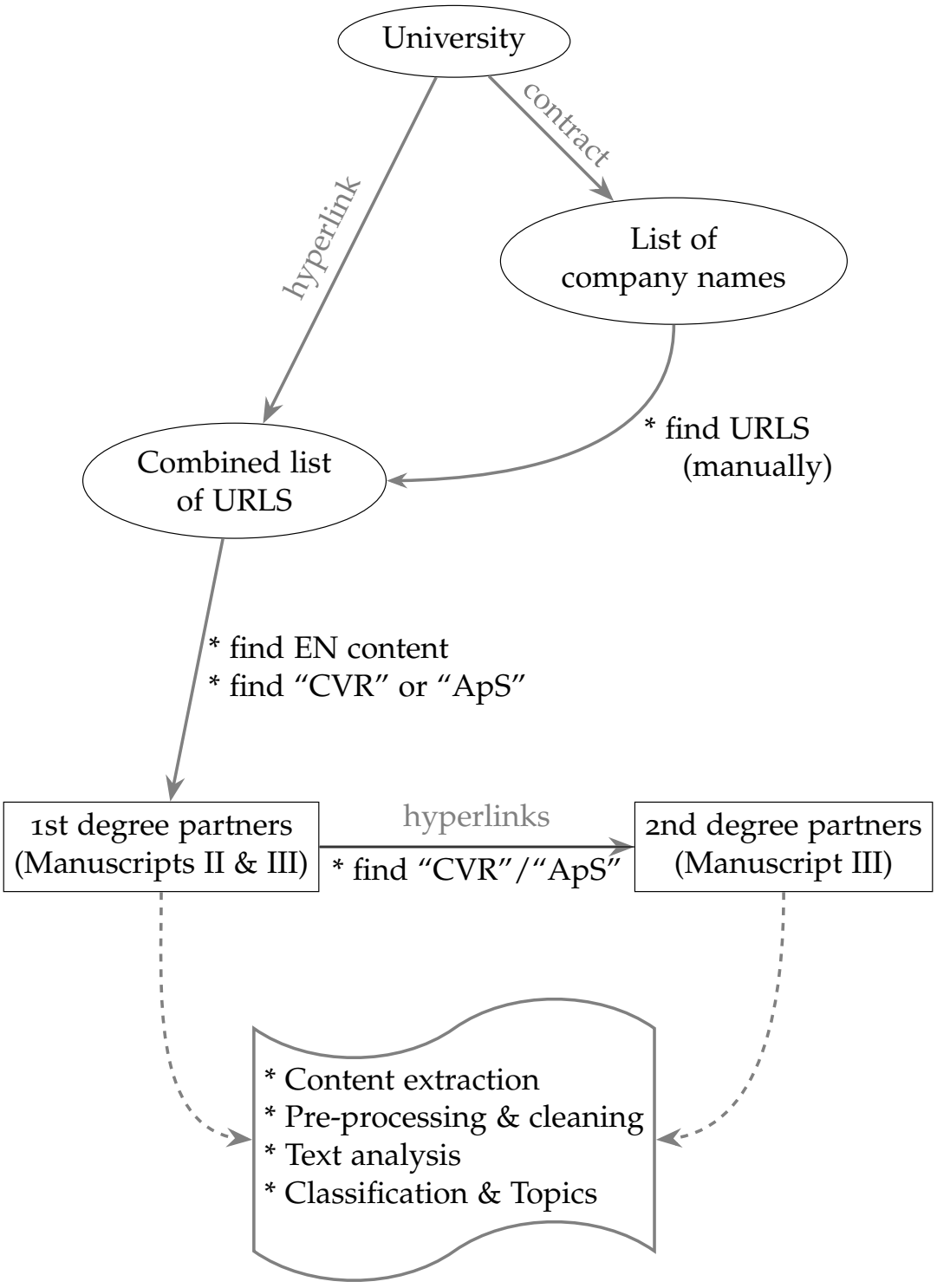


Figure 8: Overview over the generation of the two main company data samples.

remaining entries might be publicly funded in which case they were removed in the later cleaning steps. The HTML content was collected from the identified websites and stored in plain text. Of the 1402 entities 1221 could be retrieved online. The text content gathered was used to identify the above-mentioned requirements. First, each piece of content was checked for the language parameter and only English texts were retained. Second, identification of Danish registration was performed (on all the retrieved websites) via regular expressions and pattern detection (for more details see Manuscript II) and A).

The detection had to be verified manually which exercise revealed only a few false positives meaning companies with no Danish registration but which had been retained in the sample. This referred to cases where a date or telephone number had the same pattern as a Danish VTA number (CVR) number. (Almost the same cleaning procedure was applied to the second degree partners but without considering Danish registration).

First degree partners included 681 companies identified with a Danish registration. Of these only 445 fulfilled the size requirement and were used in the final sample check. More than 6,000,000 web-pages including all available PDF or similar text formats were retrieved. For the second degree partners more than 48,000 potential websites were identified from which more than 28,000 were collected. To use these for comparison in the Manuscript III. I randomly selected 700 companies to compare their output to that of the 1st degree partners.

Initially, I considered using only the Danish registered companies in the second degree sample, however, the overlap between 1st degree and second degree partners was too great to allow this approach (from around 700 companies more than 400 overlapped). The number of topics related to the first degree websites corpus was set to 45, assuming a minimum of 10 pages represented one topic. The general distribution of terms (T) and pages (P) can be found in Table 2, which shows that there are only very few very large websites that might determine the sample. The table shows also the mean and the median (Med) to provide better clarity.

P	P Mean	P Med	Terms	T Mean	T Med
138544	311	69	2185191	4911	2233

Table 2: Page and term numbers per website (first and second degree combined)

I used topic modelling to classify the companies (see Manuscript II) and to understand the basic content represented in the samples. Given the amount of data and the desired outcome some additional pre-processing and cleaning was needed to ensure a feasible approach: The identification of topics from company websites used one corpus of content from all relevant websites. However, in the absence of additional cleaning company- and product names and similar specific terms drive the detection algorithm and generate barely usable topics. Hence, additional cleaning of the word content seemed appropriate. Using an English dictionary containing more than 20,000 words (including scientific words) to restrict the content to common English facilitated this task and improved the outcome. There was a pattern among 1st degree companies showing that health, energy and services were the main areas of interest (see Manuscript II and Table 3). The second degree companies had similar features and therefore were suitable as a supplementary sample (Manuscript III).

Table 3: Example topics for the company websites with their top terms-1st degree partners

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
design	gas	hear	health	product	share
product	oil	loss	sustain	food	report
partner	develop	implant	board	process	annual
custom	report	support	report	sugar	cash
read	million	sound	news	farm	market
Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
water	drink	lab	network	oil	health
power	milk	cell	support	vessel	journal
plant	cream	order	data	gas	research
system	process	center	center	ship	clinic
pump	fill	support	switch	power	medic

8 MANUSCRIPT II

This Manuscript is licensed content from the publisher *Springer Nature*, in the journal *Scientometrics*. License obtained under the *license number: 4475291456666* for the purpose of this thesis. This Manuscript displays the findings and results for the RO2a and RO2b of this thesis.

Scientometrics (2018) 117:449–472
https://doi.org/10.1007/s11192-018-2849-9

Tracing university–industry knowledge transfer through a text mining approach

Sabrina L. Woltmann[1] · Lars Alkærsig[1]

Received: 9 March 2018 / Published online: 23 July 2018

Abstract
This study investigates knowledge transfer of university research to industry moving forward from traditional indicators by using methods from computational linguistics. We introduce a novel empirical use of pattern recognition and text mining tools to compare scientific publications to company documents. The contribution of the paper is twofold; first, a new method for tracing knowledge transfer is suggested and, second, our understanding of university–industry knowledge transfer is increased by introducing an additional perspective. We find that common text mining tools are suitable to identify concrete chunks of research knowledge within the collaborating industry. The method proves direct links between published university research and the information disclosed by companies in their websites and documents. We offer an extension to commonly used concepts, which rely either on qualitative case studies or the assessment of commercial indicators for the assessment of university research. Our empirical evidence shows that knowledge exchange can be detected with this approach, and, given some additions in the tools selection and adaption, it has the potential to become a supplementary method for the research community.

Keywords University–industry collaboration · Knowledge transfer · Text mining

Introduction

Impact is of increasing importance for universities in addition to traditional tasks of research and teaching. This Third Mission means that many universities expand their efforts beyond the production of knowledge to translate it into socioeconomic relevant contributions (D'Este and Patel 2007; Etzkowitz et al. 2000). Driven by the need to make their research known in order to secure (public) funding, universities implement various forms of transfer activities, such as adaption of strategic licensing and university patenting, which ensure that their findings are (commercially) utilized (Gulbrandsen and Slipersaeter 2007). However, the detection of relevant knowledge transfer remains non trivial. Thus,

✉ Sabrina L. Woltmann
swol@dtu.dk

[1] Department of Management Engineering, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark

new methodologies are needed to quantitatively assess knowledge transfer of universities to enable a more holistic analysis. This paper examines the transfer of university research to the industry through a novel combination of methods based on established text mining tools. We use additional data sources and metrics, to move beyond the traditional proxy indicators. We aim to identify the potential and limits of contemporary text mining tools for a detection of identical knowledge pieces. Text documents are already used in numerous studies as data sources and are, hence, suitable to answer relevant present-day questions (Zhang et al. 2016). Our approach is unique, since it captures identical knowledge pieces in the university and the industry in its (geographical) proximity. The intention is to capture the transfer without focusing on specific transfer channels, collaboration types or related commercialization mechanisms. The knowledge detection is made through the application of existing text mining methods, namely the latent Dirichlet allocation (LDA) and the algebraic indexing method called term-frequency, inverse document frequency (TFIDF). LDA is a known topic model, used to identify underlying structures in entire text collections, while TFIDF indexing can be used to extract keywords for single documents. We use a combination of both methods to identify *identical* knowledge pieces.

Literature

Knowledge transfer concerns "(...) the conveyance of knowledge from one place, person or ownership to another. Successful knowledge transfer means that transfer results in successful creation and application of knowledge in organizations" (Liyanage et al. 2009, p. 122), including the necessity of utilization of this particular knowledge.[1] Given the particular role of universities within the field of knowledge transfer, a great deal of literature has established a well developed empirical basis for the assessment of university driven knowledge transfer (Agrawal 2001; Perkmann and Walsh 2007). The empirical approaches are often derived from integrated models on the institutional level, such as the triple helix model (Etzkowitz and Leydesdorff 2000a), which are regularly reduced to a bilateral university–industry focused concept that investigates collaborations between universities and firms on individual, organizational or national level (Siegel et al. 2003; D'Este and Patel 2007).

Additionally, the research is also divided into formal? and informal knowledge transfer. Formal transfer will eventually "result in a legal instrumentality such as, for example, a patent, license or royalty agreement (...)" (Arundel and Marcó 2008, p. 642), while informal transfer is seen as resulting from informal communication and does not lead to outcomes that fall under intellectual property regulations (Tijssen et al. 2009; Link et al. 2007). Overall, the main attention in the literature on university–industry knowledge transfer has been given to formal knowledge transfer often focusing on the commercial value the knowledge yields (Thursby et al. 2001; Wu et al. 2015; Han 2017). Due to the abstract nature of knowledge transfer its actual measurement remains challenging and relies heavily on proxy indicators. The indicators for formal (commercialized) knowledge transfer, even though well developed, fail to measure instances where knowledge cannot easily be commercialized, patented or licensed (Cheah 2016; Cohen et al. 2002; Agrawal and Henderson 2002). These circumstances have left the research community with a gap in

[1] Technology transfer and knowledge transfer are in the literature strongly interrelated concepts and are widely used as interchangeable terms (Grimpe and Hussinger 2013; Sung and Gibson 2000).

tracing and measuring the university–industry knowledge transfer (Sung and Gibson 2000) and the need to investigate and assess potential new methods.

Text mining: empirical applications

One of the contemporary approaches to solve various kinds of measurement or detection challenges is the application of data mining or in particular text mining (Aggarwal and Zhai 2012). In this regard, computational linguistics, the scientific base of text mining, became increasingly relevant for empirical studies in a number of unrelated academic fields (Yau et al. 2014; Aggarwal and Zhai 2012; Gaikwad et al. 2014). Previously great insights in disciplines like social sciences, biology, and economics have been achieved through the use of text mining tools (Zhang et al. 2016; Garechana et al. 2017). Text mining applications have also gained traction within research concerning knowledge networks and knowledge flows (Magerman et al. 2010; Leydesdorff 2004). In studies investigating the influence on knowledge generation and dissemination of universities, the triple helix model (Etzkowitz and Leydesdorff 2000b) is often used as a foundation to unveil concrete knowledge linkages (Meyer et al. 2003). These studies aim to measure the underlying structures of the (knowledge driven) relationships between governments, academia, and industry and apply regularly text mining based measurements (Khan and Park 2011). These contemporary text mining applications are often used in combination with other bibliometric tools. An evaluation of university–industry interaction can, for instance be done through and the identification of key words and co-occurrences (Khan and Park 2011). So today's understanding of the triple helix interaction has been immensely increased by relying on bibliometric, network and text mining approaches (Glänzel and Thijs 2012; Zhang et al. 2014).

However, approaches on tracing knowledge transfer remain at a relatively rudimentary level. Even though some studies show the successful application of text mining methods (Van Eck and Waltman 2017; Tussen et al. 2000), the concrete outcomes remain often undetected. The application of these methods also remain challenging (Meyer et al. 2003). The main challenges today include the identification of the actual contributions of university research without limiting analyses to too narrow indicators or being to imprecise. Often only trends are detected, since measures like citations and references do not hold up well in an industry context (Jaffe et al. 2000). Hence, new detailed measurements for knowledge transfer are needed. Our study provides an assessment of the use of text mining methods to extract relevant pieces of knowledge from universities and identify them within companies' public documents. The contribution and innovative approach of this study is to identify the concrete pieces research, such as the results of an experiment or a novel method, from a university publication base and trace them.

Methodology

We focus on knowledge transfer overall, which particularly includes the aspects of technology transfer.

Our approach is different to conventional knowledge flow detection in the sense that we aim to identify concrete research outcomes including for instance a concrete technology, method, algorithm, chemical formula etc. and focus less on similar working fields or just coherent topics. This focus makes the actual identification and verification more challenging in a technical sense. We use a combination of two well known techniques the latent

dirichlet allocation (LDA) and the term-frequency, inverse document frequency (TFIDF). This combination allows the extraction of relevant keywords per text and also for entire text collections, which allow a keyword comparison.

Generally, text mining can be used to describe the extraction of knowledge from free or unstructured text. This encompasses everything from information retrieval to text classification and clustering (Kao and Poteet 2007). Current rapid developments in computational linguistics provide improved accuracy and feasibility (Chapman and Hall/CRC 2010; Collobert et al. 2011). The task of identifying content similarity, however, remains up until today challenging. In particular as the similarity between linguistically highly diverse texts remains widely unsolved.

In the following, we outline the particular methods and algorithms used to fulfill the study's objectives. We aim to give insights into current developments in the field as well as determine the used methods and specify, parameters and tools used. We aim to trace concrete research outcomes from the university, which requires key word extraction, comparison, efficient pattern recognition and similarity measures.

Identifying similarities within texts is a very particular area of text mining. Similarity measures can be based on probabilistic as well as algebraic models. However, these practices are often used to detect actual paraphrasing and these models are limited to identify word-to-word or phrase-to-phrase similarity (Rus et al. 2013). However, for our purpose these applications are too narrow and focus plainly on the linguistic composition and are only applicable on extremely short text snippets.

We make use some of the same basic tools, but combine them in different manners to identify overlapping content for entire documents (Fig. 1).

As we aim at tracing concrete research results from the university, it is necessary to combine the comparison between topics and the TFIDF indexing. Therefore, it is not enough to identify that two corpora (a website and a department) share the same topic, for instance 'wind energy', but that for instance a new assessment model (developed and described in the publication) is used by the company. This insight can only be generated on a document-to-document level, but needs to be supported on a corpus topic level. This is crucial, since there are no other concrete indications for transfer, such as citations or references.

Pre-processing

To apply text mining procedures, the pre-processing of the data is essential. It entails data cleaning and additionally conversion of unstructured raw text into statistical and computational useful units. The quality of text mining results is highly depending on the thoroughness of the pre-processing. The main objective is to capture relevant characters and erase obsolete items (Paukkeri and Honkela 2010). We follow the procedures as described by Ponweiser (2012, p 33), i.e.:

- Define word boundaries as white spaces,
- Delete unwanted elements (e.g. special characters, punctuation, ...),
- Convert all characters to lower case,
- Remove stopwords (common words that don't carry content information),
- 'Stemming' words, this reduces words to their morphological word stem (Schmidtler and Amtrup 2007, 126),
- Remove words that are shorter than three characters.

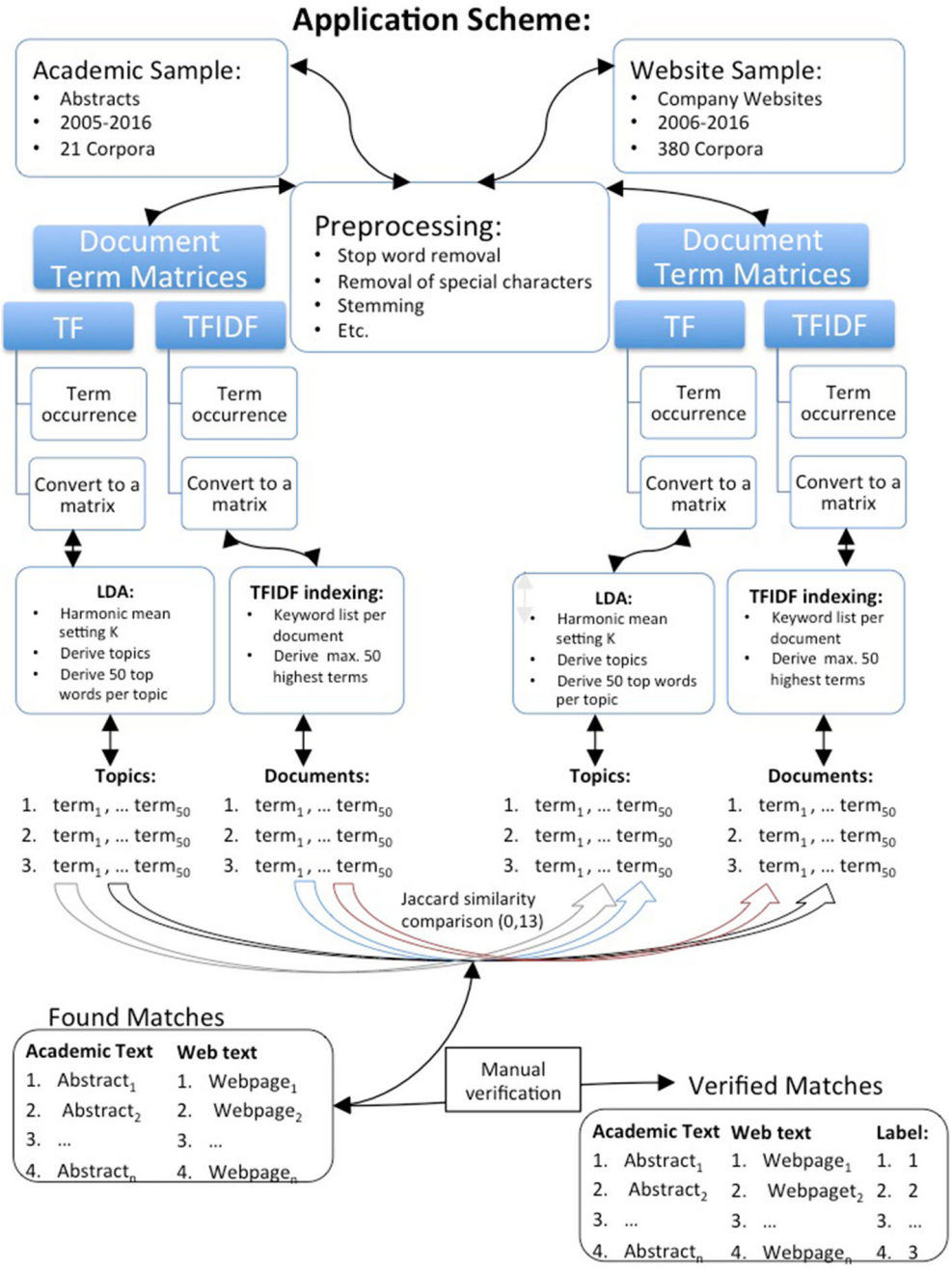


Fig. 1 Steps of statistical method application for the different samples

The pre-processed texts are merged into structured units and, in our case, also thematically classified units, the *text corpora*. To prepare the texts collections into a statistically useful format, the corpora are converted into *document-term matrices*. A document-term matrix is the most common vector space representation of document corpora. Rows correspond to documents and columns to terms. It contains the feature (term) frequencies (number of occurrences) for each document (Richardson et al. 2014; Chapman and Hall/

CRC 2010). These matrices are usually highly dimensional and sparse and accordingly most text mining methods most include dimensionality reduction (Berry and Castellanos 2007).

In a document-term matrix the element at (m, n) is the word count (frequency) of the i'th word (w) in the j'th document (d).

$$Term - Documentmatrix = \begin{array}{c} \\ d_m \end{array} \begin{array}{c} \begin{array}{cccc} & & w_n & \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,n} \\ \vdots & & \ddots & \\ x_{m,1} & x_{m,2} & \ldots & x_{m,n} \end{pmatrix} \end{array} \tag{1}$$

Term weighting and indexing schemes

Various *term weighting schemes*, determining the value of each entry, are available. The weight for each term can be derived by the application of different measures and is based on the frequencies of term occurrences. Specific text mining models rely on a particular term weighting input (Xia and Chai 2011).

- Binary weighting takes values 1 or 0 depending on whether or not a term occurs,
- Term-frequency (TF), which is the actual number of occurrences of a term for a given document.
- Term-frequency, inverse document frequency (TFIDF), assigns higher weight to terms that occur in a small number of documents (Xia and Chai 2011).

The TFIDF is a simple numerical indexing method, which has been applied in various contexts (Franceschini et al. 2016; Zhang et al. 2016) and gives respectable results on its own, but it also serves as basis for various more advanced models, like the Vector Space Model (VSM) or Latent Semantic Analysis (LSA) (Mao and Chu 2007).

The principal assumption behind the TFIDF is that words that occur often in a document are relevant for its content, but words that are used in many documents are less content specific for the single document. Frequent words that are used in many texts carry less contextual information and obtain a lower score (Robertson 2004). TFIDF indexing enables a dimensionality reduction providing a small set of content relevant terms. Most commonly the TFIDF is calculated by multiplying the term frequency *TF*, the number of times word w appears in document d; and the inverse document frequency *IDF*, which is the logarithm of the total number of documents D divided by the number of documents that contain the word w denote dw (Aizawa 2003).

$$\begin{aligned} \mathrm{TF}(w,d) &= \sum w_i \\ \mathrm{IDF}(w,D) &= \log\left(\frac{D}{dw}\right) \\ \mathrm{TFIDF} &= \mathrm{tf}(w,d) \times \mathrm{idf}(w,D) \end{aligned}$$

The TFIDF approach suffers from some shortcomings. First, it might represent only the content of a particular text fragment, which is a major drawback for long texts. Second, IDF assumes that terms, which rarely occur over a collection of documents, are more content related, while in reality they are just more distinctive. Third, empty terms and

function terms are often assigned too high scores (Xia and Chai 2011). Nevertheless, the TFIDF approach has been proven to provide very robust and high quality results (Robertson 2004).

For the purpose of this study, we use (among other metrics) the TFIDF indexing to determine the most characteristic words for each document. Hereby we reduce the dimensionality and enable a comparison of keyword of different texts with each other. Hence, the lists, generated for each document are used to identify common terms between two types of documents, abstracts and website pages.

Latent Dirichlet allocation (LDA)

LDA is an application of topic modeling and is a fully automated method based on statistical learning, which aims to identify latent (unobservable) topical structure in a text corpus (Blei et al. 2003; Griffiths and Steyvers 2004). LDA extracts underlying structures of texts and translates them into topics, which are composed of terms that are assigned together with a certain probability to each topic.

LDA works as follows, described by Grün and Hornik (2011, p. 4) and Ponweiser (2012, p. 15):

1. For each topic, we decide what words are likely (term distribution described as $\beta \sim \text{Dirichlet}(\delta)$
2. For each document,
   (a) we decide what proportions of topics should be in the document, (topic proportions defined by $\theta \sim \text{Dirichlet}(\alpha)$.
      i. for each word in the document:
         A. we choose a topic $(z_i \sim \text{Multinomial}(\theta))$.
         B. given this topic, we choose a likely word (generated in step 1.) from a multinomial probability distribution conditioned on the topic $z_i : p(w_i|z_i, \beta)$.

To improve the performance of the LDA we added one pre-processing step that excluded terms, which occur in more than 90% of the documents in the document-term matrix. The resulting topics are more specified and do not contain generic terms. The LDA algorithm needs to start with a pre-defined number of topics denoted K. Separate approaches were used for estimating K for the academic corpora and for the companies website corpora. For the academic abstracts, K was estimated using the following approach: we approximate the marginal corpus likelihood (depending on K) by taking the harmonic mean for each corpus after applying LDA for different numbers of K. Hereby we are sampling the best 'fit' for a set of possible K values. The harmonic mean takes one chain of samples as argument to first collect all sample log-likelihoods and subsequently calculates the harmonic mean of these likelihoods. The log-likelihood values are determined by first fitting the model and to do this over a sequence of topic models with different numbers of topics. This is an approximation of $p(w|K)$, i.e., the likelihood of the corpus given the number of topics (Ponweiser 2012). The upper level for K was set to 200. However, this method is computationally very expensive and is therefore only feasible for the shorter texts in the academic corpora.

For the websites corpora, we set the topic number according to each individual corpus size. We simply use the total number of documents for setting K, assuming that a larger corpus contains more distinct topics:

$$D_m \geq 3000 : K = 200$$
$$D_m \geq 2000 : K = 150$$
$$D_m \geq 1000 : K = 100$$
$$D_m \leq 1000 : K = 50$$

The hyper-parameters for the LDA are in our case aligned to the needs to identify common content rather than to classify a document into a topic. Hence, we use the Gibbs sampling for determining the posterior probability of the latent variables. We use standard $\alpha = 50/\mathrm{k}$ as parameters of the prior distributions. For more information on determining the posterior probability of the latent variable see Grün and Hornik (2011).

Jaccard similarity coefficient

To measure the similarity between the sets of identified keywords, we use the Jaccard similarity coefficient as metric (Niwattanakul et al. 2013). We chose this similarity measure as it only includes element presence in a given set. It is applicable for the LDA and/or TFIDF generated keywords. This has two major advantages for our purpose: First, Jaccard similarity is not based on the input of scores or probabilities, which would in our case be hard to compare, since they result from different corpora and even usual normalization's are not necessarily good enough. Second, the overlap over terms is comparatively low, due to the high linguistic difference between academic writing and public websites, which is not the case for most other studies, focusing on more similar types of documents (Zhang et al. 2016). Therefore in our case a set comparison is more relevant. The similarity measure yields scores that are highly dependant on pre-processing and data type, and therefore needs specifically adjusted thresholds for our study. However, this said, it is not given that in other circumstances with similar goals other similarity measures, such as the cosine similarity or euclidean distance will not be more appropriate.

The Jaccard similarity is based on the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, 1 indicating most similarity (identical sets) and 0 indicating least similar: no common feature in the two sets. Given the set of keywords from one document of the publication database denoted K_A and the second set of keywords from one page of the websites denoted K_B, the Jaccard similarity denoted $J(K_A, K_B)$ is obtained with:

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A| + |K_B| - |K_A \cap K_B|}$$

The thresholds for a minimum similarity for further examination were chosen based on preliminary results. In all applications, we only consider it a potential match if keyword lists return a certain minimum Jaccard similarity. However, the Jaccard similarity tends to benefit smaller sets. Hence, we decided to set a common threshold to a minimum of 0.13 and another used indicator threshold consisted in multiplying the Jaccard similarity with the intersection of the two sets, giving higher weight to sets with a higher amount of common words. Two sets with Jaccard similarity lower than 0.15 need more than 7 words

in common in order to pass the criteria, while set pairs with Jaccard similarity higher than 0.15 can have smaller intersections.

Manual classification and verification

To determine whether the findings of the algorithms are actually relevant or valid, we needed human inspection and final verification. This is necessary since we are working with unlabeled text data and would not be able to verify the results without human confirmation. This step verifies the data and enables insights about the performance of the computational tools. Especially, since the data sample does not provide the possibility to have a labeled training data set, meaning that we have no training data, which could offer objective labels for the text matches. However, this is not really a possibility due to the huge amount of text pairs and the high level of complexity of the text documents.

We used three independent people from different disciplines to decide about the similarity of the text snippets. They were asked to categorize the text matches into one of five categories:

1. Identical topic = University contribution
2. Identical topic = Potential university contribution
3. Common topic = Unlikely directly related
4. Different topic = No match in content
5. Unclear = could not be classified

In the first label we included also findings about identical topics, which are a University contribution, but to a public entity, or media article or news about university research. If needed the people could resort to the actual full text publication, in case the abstract did not provide enough information for a final verification.

The human result classifiers background is as follow: three academics (Ph.D. students) from different fields and one engineer. A fourth person was then making final decisions when disagreement is observed between the three human classifiers. The general idea is to use people that are capable to identify research topics and applications in various context.

Test data sample

To test our text mining methods we use Technical University of Denmark (DTU) and its economic environment as example case. To establish a first test sample, DTU is an appropriate case for this research for the following reasons: First, focusing on a technical university enabled us to study leading edge technology research with direct connections to industry innovation. Second, DTU provides a well documented case and the number of research institutions in Denmark is rather small, which allows straightforward attributions to a specific university. Third, Denmark has a high level of digitization and data availability, making it a promising setting for applying text mining. The scope is ideal as a first use case especially since DTU has already a comparatively high level of commercially relevant knowledge (http://www.dtu.dk/english/Collaboration/Industrial_Collaboration) and industry ties, which supports the assumption that there it is a fruitful case for tracing knowledge transfer. The type of research is very applied and hence highly relevant to the private economy.

As we aim to detect knowledge transfer from universities to the industry, we use the research output of the university as baseline since publication texts are the formalized

output and dissemination channel of university research and contain all important research findings of a university (Toutkoushian et al. 2003). On the other hand, we use websites, which are companies channels used to ensure their visibility for potential consumers and investors including their most recent R&D successes and collaboration efforts (Branstetter 2006; Heinze and Hu 2006). The comparison of these two sources aims to detect knowledge overlap seems feasible.

Furthermore, Denmark, as national context, is ideal as its research is almost exclusively published in English language and most companies also use English as secondary, if not as first corporate language. This is highly relevant for the application of the text pattern recognition and for co-word occurrence measures.

Publication database

We focus only on recent research outcomes by a university and exclude widely known and commonly accepted knowledge. Therefore, only novel scientific insights, technological innovations, like leading edge technologies shape the scope of this study.

To identify relevant university research, we use the universities publications published by the university between 2005 and March 2016 . In the case of DTU, the data is taken from a database named ORBIT http://orbit.dtu.dk/en/. The retrieved entries present main research outputs by at least one employee of the university. However, the registration of research items only became mandatory in the year 2012, so it is important to mention that data coverage is not equal across the all years of the observation. The data provided by the database include a collection of academic abstracts, open-source full-text publications and publication meta-data. The meta-data includes among others: year, author(s), title, journal name, university section id, internal id, and DOI (digital object identifier). The number of all publication records for the time period is 78,466. For more detailed information on the available publication data, see Table 1. We cleaned the abstract data by removing all entries, which had no real text in the abstracts field, which resulted in 55 removed entries.

We classified the texts (abstracts and full-texts) by their database assigned departmental codes, which we converted into collections of research areas. This provides a pre-classification of texts by their fields. The sub-setting resulted in 24 separate research fields (see Sect. 2) of which three are irrelevant for the academic output of the university. (We excluded approximately 250 articles including (1) publications registered to the university administration, (2) publications registered to the bachelor program, and (3) one set that was directly linked to a large company). The collection of these research area based corpora will in the following be referred to as ‘academic’ corpora or by their individual name if this is relevant for the interpretation of the results. Most text mining methods perform better on more contextual coherent corpora and hence achieve better performances.

Table 1 Total publications for the years: 2005–2016

Year	Abstracts	Only texts	Abstract or text	All publications
2005–2010	16,502	2738	3854	40,455
2011–2016	28,517	5137	11,963	38,011
2005–2016	45,019	7875	15,817	78,466

The distribution shows that the coverage and also the research output varies a lot between the research fields. This is crucial to keep in mind when analyzing the amount of observed knowledge transfer according to the fields. Especially, given that fields like Nuclear technology have only 316 abstracts but a high coverage since the entire output is only 422 articles, this might be due to the size of the research group at the university and/or the groups age (see Sect. 2)

We chose the abstracts to serve as main research sample. This shortens computational time and enables better investigation of relevant fields and texts. The findings from this preliminary analysis are then used to find most relevant corpora for more in-depth and more extensive exploitation of the methods.

Companies

The second data source, providing the company knowledge, was gathered from corporate company websites, since knowledge chunks, which are displayed on a website have to be of a certain commercial relevance for a firm.

First, we identified the key criteria for relevant companies, which are defined as: (a) having a national (Danish) company registry number (CVR) and (b) having had a collaboration contract with the university between 2006 and 2016. This constitutes a direct formal link between the companies and the university, which is the ideal basis to test and verify the new method.

To identify more potentially relevant companies, we generated one network on the basis of hyperlinks between the university and company websites. Hereby we identified additional partners linked to the university website. The list of websites contained many online service platforms. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample.

The websites themselves needed to provide as a minimum a set of 5 English web-pages with in English minimum of and more than 100 English words per page and display the CVR number on the website. We fetched the HTML content of the websites using a self designed web-crawler (https://github.com/nobriot/web_explorer) and converted it to usable plain text cleaning it from any remaining code tags. These online text samples were collected between August 2016 and November 2016. Exploring the websites, we visited 908,288 total web-pages (single text documents in total), that had to be filtered by the above mentioned criteria for websites.

The number of total number of companies, which could be identified as collaborators of the university between 2006 and 2016 was 1225 of which 699 had a CVR number written on their website and 544 were displaying the Anpartsselskab (ApS) abbreviation (which describes limited liability companies in Denmark). Certain companies went out of business, underwent mergers or were just renamed. We tried to identify the new names or entities, however this was not in all cases possible. We were left with a final sample of 445 companies. The firms in this sample operate mainly in technology intensive sectors and are firms with strong R&D divisions. Therefore it included companies with contents related to the research performed at the university.

To provide an overview of the composition of the firm sample we decided to identify the main industry field of each company by using additional text based tools. This is reasonable since the identification of topics and clustering of texts has a long tradition and has successfully been used in various research areas.

We applied the LDA for the clustering of companies (for more details see Sect. 3 to identify the main categories for the firms, showing the overall distribution of firms that work within one topic or field. We used our knowledge of the sample to set the optimal number of topics ($K = 45$). To avoid too generic topic clusters we erased all words that were used in more than 80% of the websites, which removes website specific terminologies, such as the contact information, impressums and similar. For a better understanding we summarized the single topics with their most relevant keywords for each topic (see Table 2). The clustering cannot be assumed as reliable as the labels from the scientific fields, however they show clear focus in some fields (see Table 3).

The number and length of pages varies a great deal between company websites (see Table 4). Some have an English summary for their main contents, while others, often multinationals have their entire website in English. This difference in length clearly influences the performance of the statistical models, since long text documents generally influence these models more than short ones. In this sample collection, we also ensured to capture the content of PDFs or similar formats stored on the websites. These required special treatment and are treated as pages of the websites. Each website is stored as its own corpus. Even though this might seem drastic, it is a sufficient way to ensure a comparable pre-classification like research fields and fosters the performance of the statistical methods.

Results

We divide this section according to the results of each applied method to give an explicit insight into the performances and future potential of the single applications. It is crucial to keep in mind that this study is a first step to verify effectiveness, limitations and eventually identify applicable thresholds and suggest future improvements. Finally, we set the results into context and evaluate the outcomes based on the studies objectives. In each subsection we clearly describe which data samples are used and why. This is crucial because of the varying demands of the different methods. The different methods generated different outcomes in terms of keyword lists, due to their different levels of application (document or corpus level) (see Table 5).

Our pre-processing revealed some specific challenges, in particular in the case of the academic abstracts. The abstracts contain, for instance, chemical formulas and notations, which rely heavily on numbers and/or special characters. These are removed during the course of the pre-processing and therefore lost in the subsequent application. The only possibility to later identify same formulas to use them for similarity measures is the assumption that the removal of those characters will always result in an identical end character string, but it might not always be the case. Often the result may not be identifiable as the particular formula, but still provides a match. In some rare cases HTML, or other code tags prevented the identical deconstruction and in such cases, we did not find a way to identify the matching strings. However, some terms may seem like the result of poor pre-processing, but are in reality just a representation of specific models, formulas or project names shrunk to an unidentifiable string of characters. The websites on the other hand are challenging in a different way: they contain different language snippets, which are embedded in every site forcing language detection on lower levels. Therefore we decided to only integrate web-pages that have a minimum of 80% English terms. Additionally we found that the linguistic composition of websites is comparatively repetitive within a website, meaning that the words companies use to describe products or services are not very diverse, which leads to high number counts for single terms. Publications, on the other

Table 2 Example topics for the company websites with their top terms

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Design	Gas	Hear	Health	Product	Share
Product	Oil	Loss	Sustain	Food	Report
Partner	Develop	Implant	Board	Process	Annual
Custom	Report	Support	Report	Sugar	Cash
Read	Million	Sound	News	Farm	Market
Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
Water	Drink	Lab	Network	Oil	Health
Power	Milk	Cell	Support	Vessel	Journal
Plant	Cream	Order	Data	Gas	Research
System	Process	Center	Center	Ship	Clinic
Pump	Fill	Support	Switch	Power	Medic
Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18
Custom	Drill	Wind	Light	Cancer	Plan
Data	Reservoir	Project	Electron	Influenza	Consult
Platform	Seismic	System	Power	Prevent	Project
Network	Fluid	Public	Wire	Flu	Design
Cloud	Data	Product	Tool	Control	Environment

hand, have a much richer vocabulary and therefore suffer less from this skewed word distribution. To account for this different composition of the two text types we normalized or removed the words in question when needed.

Text comparisons

To identify potential text documents with identical knowledge pieces we first compare the keywords from publications and websites with the computational methods. Hereby, we identify text pairs that potentially contain identical knowledge content. However, in the final step these potential matches have to be manually verified.

The keywords are derived through TFIDF indexing or extracted from the topics of LDA. The LDA on the academic corpora resulted in 915 distinctive topics for all 21 academic corpora. The LDA for the websites resulted in 8250 distinct topics (see Table 5).

To verify the performance of the LDA application, we manually inspected the derived topics for several corpora to ensure the performance, including the decision regarding topic numbers and prior settings.

The manual inspection revealed a much clearer picture with the academic texts than with the websites. The topics for the single scientific areas seem very distinct and reasonable (see Table 6).

While the LDA applied to websites still gave some good indication about their main area, the topics seem less distinct. However, the manual inspection suggests that LDA is capable to represent the main content of a website, but due to the previously mentioned word repetition adjustments in terms of too frequent words need to be made. Accordingly, we removed all terms occurring in more than 90% of documents in a website.

Table 3 Topic distribution of the websites

Topic no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	26	3	4	2	37	6	1	9	12	7	1	7	21	25	1
Topic no.	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	34	1	16	23	13	9	2	2	3	2	7	2	10	2	7
Topic no.	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	1	61	2	17	5	4	2	11	1	3	21	3	4	13	2

Table 4 Page and term numbers per website (descriptive)

	Pages (P.)	P. Mean	P. Median	Terms	T. Mean	T. Median
Total	138,544	311	69	2,185,191	4911	2233
Lower boundry	5	–	–	38	–	–
1st quantile	22	12	10	905	521	523
2nd quantile	69	42	40	2233	1476	1408
3rd quantile	257	142	130	6018	3819	3675
4th quantile	10,106	155	591	67,351	13,866	10,466

Table 5 Keyword lists

Method	Total number of topics/keyword lists	Number of corpora
TFIDF web	138,552	380
TFIDF orbit	44,294	21
LDA topics orbit	915	21
LDA topics web	8250	340

Since the LDA can only be applied on texts that have a certain length, as the algorithm depends on the amount of text data input, we had to exclude 40 smaller website corpora. These corpora are suited for the TFIDF application, but not for LDA. Therefore the sample size of LDA is slightly smaller than for TFIDF (see Table 5). The LDA provides a certain number of topics for each corpus (these vary according to the website length (see Sect. 3.3). Each of these topics are composed of specific words, which we extract and combine to a keyword list. For the LDA comparison we selected the 50 most relevant (probable) words for each topic (LDA allows term re-occurrence in different topics with different probabilities). We compared each topic from one of the 21 academic corpora and 340 website with each other. Each time a keyword list is compared to another and the Jaccard similarity is computed for each comparison. More than 7,548,750 individual comparisons were performed.

Examining the Jaccard scores revealed that none of the comparisons scored higher than the set threshold of 0.13 (the first set threshold). The first matches between topics were around the threshold of 0.08. This is a really low similarity score and shows that the academic and web corpora are very diverse in the main areas. 12 document pairs could be

identified exceeding a Jaccard threshold of 0.08. Comparing the academic topics from different departments with each other reached scores up to 0.82 Jaccard similarity, which shows how much closer academic corpora are related.

The TFIDF provides keywords for each document, hence the number of lists equals the number of documents in each corpus. We extracted up to 50 highest indexed terms for each document (see Table 5). We compared each TFIDF keyword list from the academic documents with all keyword lists from the websites. The maximum length of the keyword lists was set to 50 extracting the words with the highest TFIDF scores (see Sect. 3). However, some texts, mostly the academic abstracts, were too short to generate a list of 50 words, hence we decided to set the length of the list of words to all words remaining after cleaning and pre-processing. We additionally excluded around ten websites, as they were too short for the application of the TFIDF. However, comparisons for shorter texts are object to the adjusted Jaccard threshold (see Sect. 3) to ensure that the short keyword lists are not dominating the final match sample with less relevant matches. We retrieved 44,294 lists for the academic abstracts. and 138,552 keyword lists for the websites resulting in 6,137,022,288 comparisons.

Compared to the LDA application some keyword list pairs scored comparatively high. 124 pairs with 0.13 Jaccard similarity threshold were identified. After a preliminary manual inspection we decided to apply another cleaning step for the TFIDF matches, since some particular matches share no contents, but only certain distinct words that are irrelevant for the content, such as foreign language fragments or country names (see Table 7). We excluded there fore all the pairs that were matched on those kind of keywords. 91 final text pairs that were after the cleaning procedures which represents a very low $1.48 \times 10^{-6}\%$. For the purpose of comparison we tested two different academic corpora from 'Mechanical Engineering' and 'Computer Science and Mathematics' and compared their TFIDF keyword lists. The assumption is that the contents are more related and the linguistic composition closer. This test resulted in 487,961,509 comparisons. By applying the same thresholds a total of 1377 matches was identified which is $2.8 \times 10^{-4}\%$ matches, way higher than in the websites against academic documents. This comparison shows that the single match between academic and website documents is more relevant, since these are not commonly coincidental. It also confirms the high diversity between the two sets of documents.

We have also compared the retrieved keywords from the TFIDF of the websites with the keywords found with the LDA topics computed on the academic corpora. We again set an upper threshold to 50 words per topic. This comparison yielded to a total of 33 matches and after the second clean-up, only 13 potential matching pairs.

To identify the actual documents belonging to an actual topic generated by the LDA is not straightforward, since only a probability distribution over documents is given. Hence, we used for each topic the two documents with the highest probability. This resulted in each TFIDF text having two potential matches for academic abstracts.

Table 6 Topic example for one academic corpus

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Enzym	Bind	Dynam	Forc	Oligosaccharid
Domain	Site	Vibrat	Particl	Branch
Amino	Conform	Motion	Hydrophob	Carbohydr
Residu	Enzym	Coupl	Friction	Donor
Express	Residu	Excit	Layer	Polysaccharid

Table 7 Example of word combinations which had to be excluded from potential matches

Countries		German		Danish	
"kingdom"	"franc"	"wird"	"auf"	"eller	"flere"
"germani"	"european"	"der"	"ein"	"som"	"til"
"poland"	"finland"	"die"	"bis"	"til"	"det"

Human verification of the text pairs

The results generated with the TFIDF to TFIDF comparison and the LDA to LDA comparisons show a significant theme overlap between the text documents. Comparing the text pairs retrieved from these both applications resulted in 10 common matches, meaning that the TFIDF and LDA returned 10 times the same text pairs. Interestingly, in the manual verification these documents are websites that achieved many hits via both applications, but refer only to overall similar content, but did not share identical research content. This means in the classical application of topics models to detect knowledge flows these pairs would have been a valid match. In our case,however, we are tracing some more specific content and these pairs do not provide clearly the same concepts, models or other knowledge. This is crucial, since these are the matches that would have been a positive identification of knowledge flows according to traditional measures using only LDA. Certain research areas revealed to be particularly dominating the text pairs, as well as in the true positives and in the entire matched sample. The overall comparison suggests a clear dominance of certain university departments in the matches. Some Departments are most represented in the matched pairs.

The combination of LDA and TFIDF reveals common interests of firms and the university and also shows which departments most are represented within the pages. Especially given that some of the comparatively small academic corpora (see Table 8) are most relevant according to the Jaccard similarity and the actual matches. In Fig. 2 we can clearly see that some departments are much more dominant when it comes to the pair-wise comparison. This means that the methods are most successful for those corpora, not determining whether it is only a content relatedness or fully identical contents. However, the other corpora seems not suited for our approach.

This is an example for a true positive, so a real text pair, which has common content and refer to the same knowledge would be the following two texts:

Academic abstract	Website document
"Swarm is the fifth Earth Explorer mission in ESAs Living Planet Programme to be launched in 2009. The objective of the Swarm mission is to provide the best ever survey of the geomagnetic field and its temporal evolution. The innovative constellation concept and a unique set of dedicated instruments will provide the necessary observations that are required to separate and model the various sources of the geomagnetic field (...)"	*"Absolute Scalar Magnetometers from CNES and CEA/LETI which were selected by the ESA for the Swarm mission. (...) The Swarm mission; a constellation of three identical satellites in three different polar orbits between 400 and 550 km altitude to measure the Earths magnetic field (...)"*

Table 8 Data coverage by research field: 2005–2017

Department	Abstract	Text	Total	% of Abst
Compute/Math	3890	1933	5791	67
Biochemistry	2343	1038	4338	54
Chemistry	1420	413	2352	60
Civil Eng.	2122	1017	3675	58
Electrical Eng.	3519	1778	4363	81
Energy Conversion	1244	521	1536	81
Environmental Eng.	1699	1269	3851	44
Management Eng.	2569	1886	4521	57
Mechanical Eng.	2999	1223	4293	70
Nanotechnology	1935	918	3064	63
Photonics	4262	2090	5617	76
Physics	1434	685	1911	75
Biology	2339	902	3562	66
Transport	860	470	1686	51
Wind Energy	1421	1158	1972	72
Food Sciences	2846	1651	6210	46
Aquatics	1481	787	4786	31
Space Research	1432	782	2137	67
Nuclear Technology	316	200	422	75
Veterinary Sciences	1520	820	2594	59
Other	2648	1954	8841	30

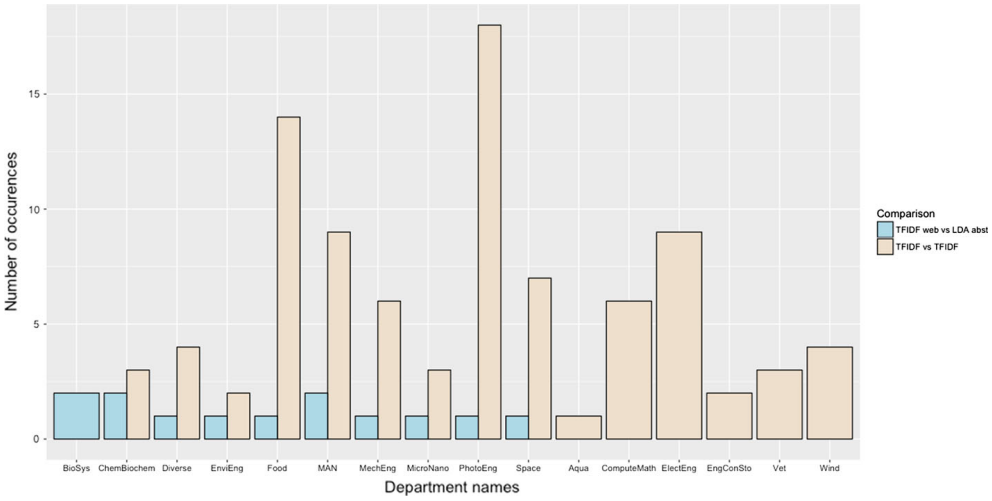


Fig. 2 TFIDF and LDA showing the most dominant research areas leading to matches that exceed the threshold

These texts show that the company is actually displaying the 'Swarm', which is the topic of the academic publication. In particularly hard cases or very limited information from the abstract the validators could fall back on the full texts of the publication.

Academic abstract	Website document
"Higher-Order ambisonics (HOA); and a matrix inversion method. HOA optimizes the reproduced sound at a sweet spot in the center of the array with radius determined by a spherical microphone array; which is used to derive the spherical harmonics decomposition of the reference sound. The four-loudspeaker-based method equalizes the magnitude response at the ears of a head and torso simulator (HATS) for sound reproduction (...)"	*"Higher-order ambisonics; matrix inversion method; ETSI TS 103 224 and matrix inversion method optimized for a specific device. For each method; the quality of the reproduced sound was evaluated both objectively and subjectively; at microphones close to a device under test and at the ears of a Head And Torso Simulator (HATS) (...)"*

The second example is according to the human verification only thematic related and does not qualify as a full match. Hence, we have to declare it a false positive. In this particular case they are very closely linked thematically, but the publication is based on the four loudspeaker method, which is not the case in the website. Therefore, these pages are labeled under category 3.

Given this examples it is obvious that the actual task is not simple and is it might appear in the first place. Therefore, we needed to ensure the quality of the assessment and ensured that several persons from different backgrounds were performing the assessment. The manual verification was performed by three persons, two researchers (PhD candidates) and one engineer, and a fourth person to handle possible mismatches in the assessment. All three are scientists and hence familiar with research and the interpretation of research results. The topic to topic comparison, with an adjusted threshold of 0.08 Jaccard similarity resulted in no positive evaluated match between texts, this confirms the assumption that the threshold has to be carefully chosen, in particular in regard to semantically very diverse texts.

Given the assessment it is clear that the engineer has a much harder time to verify identical contents that are not within his area of expertise. To see the confidence levels of each verification they made qualitative comments to their decisions, which enabled a more accurate final assessment. In Table 9 certain inconsistencies become evident. The overlap within the relevant categories 1 and 2 the low consistency was solely caused by their different understanding of the definition and was finally solved and decided based on their qualitative comments. They also commented on pairs that seemed unclear or difficult to classify to them, or in which they claimed to have specific expertise,the final labeling could be made very accurate. In particular most of the academics assessments revealed an insecurity between two labels while the other was certain about a particular label. However, the overlap for the engineer was much lower and the comments showed only a few certain classifications within his area of expertise. Revealing that mainly trained academics, used to reading academic texts, are capable to mange this tasks with sufficient confidence levels.

The fourth person (academic) had evaluate the qualitative statements, read the texts and make a final decision in alignment to the previous assessments. This strategy ensured the the quality of the results. Given the distribution of decisions (see Fig. 3) one of the main inconsistencies in the overall distribution was also the low usage of number label number 5 by the second academic, this label however should be inconsistent since it is the label for to remove the text pairs where the validator was really insecure about the labels.

The results of the verification show great overlap in content and a number of certain positive matches. As previously described the LDA comparison retrieved 12 potential

Table 9 Overlap in manual decisions

	Academic 1 & 2 (%)	Academic 1 & Engineer (%)	Academic 2 & Engineer (%)	All (%)
Total	67	61	58	48
Category 1	80	65	60	60
Category 1 & 2	74	56	50	44
Category 2	21	29	14	21
Category 3	61	54	48	38
Category 4	51	49	43	30

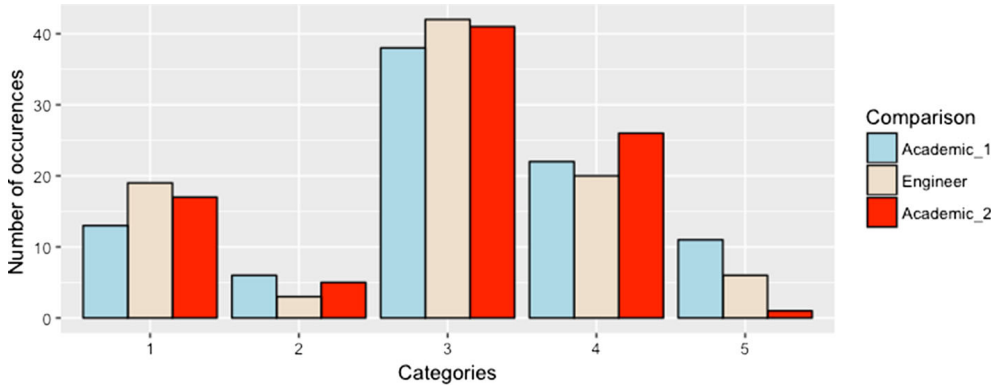


Fig. 3 Decision distribution of the manual assessment

Table 10 Number of identified potential matches

Methods	Comparisons	Matches	Matches verified
TFIDF web versus TFIDF orbit	6,137,022,288	91	20
TFIDF web versus LDA topics orbit	126,775,080	13	2
LDA topics web versus LDA topics orbit	75,487,50	12	0

matches, TFIDF 91 and the combination of both approaches 13 (see Table 10). These 12 document pairs reveled close topic connection (meaning they contained related content), but did not refer to any concrete common knowledge piece. This is not surprising since LDA is applied on a corpus level and does not in detail represent documents. Nevertheless, the common topics and themes helped to set a base for the TFIDF application. In the final verification process it was revealed that out of the 91 potential matches 27 could be verified by humans, which gives a 30% successful detection of identical knowledge pieces.

After this first comparison the performance of the TFIDF showed more success in identifying potential common contents (see Table 10). Remarkable is also that the only clearly non technical field has a high false-positive error rate.

Technical considerations

Given the progress made in the past decade the text similarity measures might become sophisticated enough to compare full texts, but for the time being we will have to apply additional strategies. For further refinement and extension, it could be considered to adopt another method for associating the documents to the LDA topics. For example, we could pick the documents connected to the highest ranking words in a given LDA topic instead of taking the highest topic probability in a document. This might be another option for future research. However, due to the size of the original sample and the complexity of the actual labeling, for now it is not possible to estimate the error on how much of the actual knowledge transfer, or the true positives are not identified.

Our findings suggest that our first estimated thresholds proved to be not accurate enough. 0.134 Jaccard distance would have been the ideal threshold for finding all text pairs for the TFIDF with a list union size close to 100 words. The best threshold would have been 0.144, here we have the best trade of between false positives and missing findings. In Fig. 2, we show the potential changes in categories (label assignment) with improved thresholds of the Jaccard measure. We lost only one match and reduced the error rates by more than 50%. The amount of first and second order matches gradually decreases with lower Jaccard similarity, as well as the content relatedness. Therefore, we suggest to evaluate the hits in future sequential, meaning to rank the hits by their Jaccard similarity and assess the first hits and stop when the amount of positive hits decreased significantly.

Conclusion

The purpose of this study was to offer new insights into both, formal and informal modes of knowledge transfer. The outcome is the development of novel detection and measurement approach for knowledge transfer, capturing instances of knowledge transfer, which are largely overlooked by current methods (Agrawal 2001). Hence, this study enables new perspectives and further in-depth understanding for reshaping existing notions on what constitutes successful university–industry collaboration in particular for policy makers and other stakeholders. It also provides generalizable and comparable findings and identifies and verifies the transfer of concrete pieces of knowledge, enabling the detection of common knowledge.

The tools we applied to detect university research as being used and displayed by private firms were indeed able to identify those instances. This study detects the use of publicly produced knowledge and moves beyond the traditional proxy indicators. Our results are not bound to the usual formal indicators and capture formal and informal knowledge transfer, as long as it is displayed from the company side. The high level of detail enables the study to show, which knowledge pieces are relevant enough for the industry to display. The trace of knowledge transfer can be directly linked to specific studies or research areas. More than 5% of the firms actually displayed some concrete knowledge driven from the university on their websites. Additionally, we still traced highly related working topics and working areas proven to be simple among the university collaborators, which adds a value to the method allowing universities to capture the most related topics with the firms in their environment (see Fig. 3 all matches contained in the third label (category 3)). In summary, the method provides insights about the transferred knowledge and is a novel quantitative assessment. It provides statistical correlation

measures, which could be used supplementary to already existing methods from the Triple Helix concept.

Even though the findings are still on a comparatively small scale, this outcome indicates that the method can successfully detect knowledge transfer. It found several instances where models, methods and clinical studies of the university were used but not directly cited. This is only a first step, but shows clearly the potential of the methods. And even though our approach reveals nothing about the underlying processes and the how of knowledge transfer from university to industry we broaden the measurement spectrum for the instances where knowledge transfer happened,regardless of the channels or mechanisms. Furthermore, the applied methods show that it is actually possible to identify concrete pieces of research knowledge in linguistically very diverse documents. This study is a first step towards an novel supplementary identification of concrete university–industry knowledge transfer.

This insight increases the understanding of the principal value of university research independently from its direct commercial success, highlighting the dissemination potential and the absorption of relevant research. Based on these findings, we might broaden the definition of 'valuable' research beyond what normally is considered valuable through patenting and licensing contracts. This would include changes within the focus on commercial value of public research, lending further support to the potential of new streams of research not identified through more traditional measurements. This could improve the funding situation for relevant but not easily commercialized research in the future, since it would enable decision makers in the university and externally to take into consideration what knowledge is actually used in the industry later. Obviously, our methods still require adjustments, but it is certainly a step to improve the understanding about public research relevance, and a strong indication that the current measures are insufficient in capturing all commercially valuable research outputs.

Future outlook and limitations

From an application perspective, several dimensions must be evaluated before the method can be widely adapted. For instance, it is crucial to benchmark the new method against the traditional indicators to assess the actual knowledge gain. Additionally, this method could be applied in different empirical settings to better understand the overall performance and application possibilities.

From a conceptual point of view, it has to be determined what this knowledge actually represents for companies and research dissemination. This estimation might not be as straightforward as it is in the case of patents or licenses, but must represent a commercial value to a company. Patents and licenses typically carry a certain commercial value, whereas the value of information on corporate websites is less understood.

From a performance perspective of the method our work can be viewed as a first step, using comparatively established methods. Technically, however, there are several improvements and bench-marking options possible. Hence, we suggest to refine the statistical methods and add more advanced statistical learning methods to improve the error rates. Focusing on the best performing research areas (see Fig. 2) would also be an option to improve the performance by strategically adjusting it to the given field.

Given these results, simpler classification might be necessary in future. Additionally, in the contrary to our expectations, the rightful classification seems to be difficult for non academics particularly when the content does not match the area of expertise. This speaks

for the high quality performance of the method: if human cannot easily distinguish false and true positives means that the method is performing well, since humans are usually performing better when it comes to this kind of tasks.

Despite the current limitations, we see clear future potential as the flexibility of the tools including potential for adaptation make them useful in various contexts.

Acknowledgements We thank the people performing the human validation of our results and the helpful comments we received on several conferences.

References

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Berlin: Springer.

Agrawal, A., & Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from MIT. *Management Science*, *48*(1), 44–60.

Agrawal, A. K. (2001). University-to-industry knowledge transfer: Literature review and unanswered questions. *International Journal of Management Reviews*, *3*(4), 285–302.

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, *39*(1), 45–65.

Arundel, A., & Marcó, C. B. (2008). *Developing internationally comparable indicators for the commercialization of publicly-funded research*. Maastricht: UNU-MERIT, 31, 1–23.

Berry, M. W., & Castellanos, M. (2007). *Survey of text mining: Clustering, classification, and retrieval* (2nd ed., p. 241). New York: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Branstetter, L. (2006). Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States. *Journal of International Economics*, *68*(2), 325–344. https://doi.org/10.1016/j.jinteco.2005.06.006.

Chapman, Hall/CRC. (2010). *Handbook of natural language processing* (2nd Ed.). https://doi.org/10.1007/978-1-4612-3426-5_15.

Cheah, S. (2016). Framework for measuring research and innovation impact. *Innovation*, *18*(2), 212–232. https://doi.org/10.1080/14479338.2016.1219230.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The influence of public research on industrial R&d. *Management Science*, *48*(1), 1–23.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning*, *12*, 2493–2537.

D'Este, P., & Patel, P. (2007). University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry? *Research Policy*, *36*(9), 1295–1313.

Etzkowitz, H., & Leydesdorff, L. (2000a). The dynamics of innovation: From national systems and mode 2 to a triple helix of university–industry–government relations. *Research Policy*, *29*(2), 109–123.

Etzkowitz, H., & Leydesdorff, L. (2000b). The dynamics of innovation: From National Systems and Mode 2 to a triple helix of university–industry–government relations. *Research Policy*, *29*(2), 109.

Etzkowitz, H., Webster, A., Gebhardt, C., & Terra, B. R. C. (2000). The future of the university and the university of the future: Evolution of ivory tower to entrepreneurial paradigm. *Research Policy*, *29*(2), 313–330.

Franceschini, S., Faria, L. G. D., & Jurowetzki, R. (2016). Unveiling scientific communities about sustainability and innovation. A bibliometric journey around sustainable terms. *Journal of Cleaner Production*, *127*, 72–83. https://doi.org/10.1016/j.jclepro.2016.03.142.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17), 42–45.

Garechana, G., Río-Belver, R., Bildosola, I., & Salvador, M. R. (2017). Effects of innovation management system standardization on firms: Evidence from text mining annual reports. *Scientometrics*, *111*(3), 1987–1999.

Glänzel, W., & Thijs, B. (2012). Using core documents for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235.

Grimpe, C., & Hussinger, K. (2013). Formal and informal knowledge and technology transfer from academia to industry: Complementarity effects and innovation performance. *Industry and Innovation*, *20*(8), 683–700.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Gulbrandsen, M., & Slipersaeter, S. (2007). The third mission and the entrepreneurial university model. In *Universities and strategic knowledge creation* (pp. 112–143).

Han, J. (2017). Technology commercialization through sustainable knowledge sharing from university–industry collaborations, with a focus on patent propensity. *Sustainability*, *9*(10), 1808.

Heinze, N., & Hu, Q. (2006). The evolution of corporate web presence: A longitudinal study of large American companies. *International Journal of Information Management*, *26*(4), 313–325. https://doi.org/10.1016/j.ijinfomgt.2006.03.008.

Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, *90*(2), 215–218.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Berlin: Springer.

Khan, G. F., & Park, H. W. (2011). Measuring the triple helix on the web: Longitudinal trends in the university–industry–government relationship in Korea. *Journal of the Association for Information Science and Technology*, *62*(12), 2443–2455.

Leydesdorff, L. (2004). The university–industry knowledge relationship: Analyzing patents and the science base of technologies. *Journal of the Association for Information Science and Technology*, *55*(11), 991–1001.

Link, A. N., Siegel, D. S., & Bozeman, B. (2007). An empirical analysis of the propensity of academics to engage in informal university technology transfer. *Industrial and Corporate Change*, *16*(4), 641–655.

Liyanage, C., Ballal, T., Elhag, T., & Li, Q. (2009). Knowledge communication and translation—A knowledge transfer model. *Journal of Knowledge Management*, *13*(3), 118–131.

Magerman, T., Van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289–306.

Mao, W., & Chu, W. W. (2007). The phrase-based vector space model for automatic retrieval of free-text medical documents. *Data and Knowledge Engineering*, *61*(1), 76–92. https://doi.org/10.1016/j.datak.2006.02.008.

Meyer, M., Siniläinen, T., & Utecht, J. T. (2003). Towards hybrid triple helix indicators: A study of university-related patents and a survey of academic inventors. *Scientometrics*, *58*(2), 321–350.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1).

Paukkeri, M. S., & Honkela, T. (2010). Likey: Unsupervised language-independent keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 162–165). Association for Computational Linguistics.

Perkmann, M., & Walsh, K. (2007). University-industry relationships and open innovation: Towards a research agenda. *International Journal of Management Reviews*, *9*(4), 259–280.

Ponweiser, M. (2012). *Latent Dirichlet allocation in R*. Ph.D. thesis.

Richardson, G. M., Bowers, J., Woodill, a J, Barr, J. R., Gawron, J. M., & Levine, Ra. (2014). Topic models: A tutorial with R. *International Journal of Semantic Computing*, *08*(01), 85–98.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*, 2004.

Rus, V., Niraula, N., & Banjade, R. (2013). *Similarity measures based on latent Dirichlet allocation* (pp. 459–470). Berlin: Springer.

Schmidtler, M. A., & Amtrup, J. W. (2007). Automatic document separation: A combination of probabilistic classification and finite-state sequence modeling. In A. Kao & S. R. Poteet (Eds.), *Natural language processing and text mining* (pp. 123–144). Berlin: Springer.

Siegel, D. S., Waldman, D. A., Atwater, L. E., & Link, A. N. (2003). Commercial knowledge transfers from universities to firms: Improving the effectiveness of university–industry collaboration. *The Journal of High Technology Management Research*, *14*(1), 111–133.

Sung, T. K., & Gibson, D. V. (2000). Knowledge and technology transfer: Levels and Key factors. In: *Proceeding of the 4th international conference on technology policy and innovation*

Thursby, J. G. J. J. G., Ra, Jensen, & Thursby, M. C. M. (2001). Objectives, characteristics and outcomes of university licensing: A survey of major US universities. *The Journal of Technology Transfer*, *26*(1), 59–72.

Tijssen, R. J., Van Leeuwen, T. N., & Van Wijk, E. (2009). Benchmarking university–industry research cooperation worldwide: Performance measurements and indicators based on co-authorship data for the world's largest universities. *Research Evaluation*, *18*(1), 13–24.

Toutkoushian, R. K., Porter, S. R., Danielson, C., & Hollis, P. R. (2003). Using publications counts to measure an institution's research productivity. *Research in Higher Education*, *44*(2), 121–148.

Tussen, R., Buter, R., & Van Leeuwen, T. N. (2000). Technological relevance of science: An assessment of citation linkages between patents and research papers. *Scientometrics*, *47*(2), 389–412.

Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, *111*(2), 1053–1070.

Wu, Y., Welch, E. W., & Huang, W. L. (2015). Commercialization of university inventions: Individual and institutional factors affecting licensing of university patents. *Technovation*, *36*, 12–25.

Xia, T., & Chai, Y. (2011). An improvement to TF-IDF: Term distribution based term weight algorithm. *Journal of Software*, *6*(3), 413–420.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786.

Zhang, Y., Zhou, X., Porter, A. L., Gomila, J. M. V., & Yan, A. (2014). Triple helix innovation in China's dye-sensitized solar cell industry: Hybrid methods with semantic triz and technology roadmapping. *Scientometrics*, *99*(1), 55–75.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, *105*, 179–191.

9 MANUSCRIPT III- WORKING PAPER

This Manuscript is partly based on the submission to the Eu-SPRI conference 2018 in Paris (see Appendix F) and the submission to the DRUID PhD Academy in Denmark 2018 (see Appendix G). This Manuscript is further developed and displays the findings and results for the RO3 of this thesis.

Open Access' influence on University-Industry Knowledge Transfer

Sabrina L. Woltmann, Lars Alkaersig, Carina Lomberg

Nobember, 2018

III.1 INTRODUCTION

Today's public universities contribute to economic development not only by providing education but also by providing access to research findings to solve social, economic, industrial and environmental challenges (Bishop et al., 2011; Kochenkova et al., 2016). Various policy measures, activities and initiatives have been implemented over the past decades to facilitate this Knowledge Transfer (D'Este and Patel, 2007; Geuna and Muscio, 2009) and access to public research has become a political focus. One of the main media used for Knowledge Transfer are academic publications. Publications can be used to measure scientific productivity and quality and to communicate scientific knowledge. They serve as tools that allow exchanges of knowledge and communication among scientists (Eysenbach, 2006), and between universities and industry (Tijssen, 2006).

For many years, access to scientific publications was only possible via subscription based services, but currently there is a trend towards open access publishing, which is supposed to promote greater use of scientific findings beyond academia and increased Knowledge Transfer. University-industry transfer of knowledge is measured mainly in terms of commercialization activities such as university patenting, spin-outs and co-publications with industry (Agrawal, 2001), with the potential implications of Open Access publications rather overlooked. There is a surprising gap related to how Open Access affects the transfer of knowledge from universities to companies (Hajjem et al., 2006; Picarra et al., 2015) compared to the extensive examination of its academic benefits (Harnad and Brody, 2004).

The present study tries to fill this gap by examining the impact of Open Access publications on certain aspects of Knowledge Transfer using a conceptual framework based on the notions of basic communication theory (Shannon, 1948) and information systems (Wilson, 2000). The research objective is to investigate the Open Access dimension and its implications for university-industry Knowledge Transfer, based on novel Knowledge Trans-

fer measures. First, we examine changes in publishing activity over the past years; second, we assess the effects of these changes on Knowledge Transfer. We focus on the changes made to the publications structure in one technical university over the course of a decade to observe developments over time. This information allows us to relate type of publication to its respective effect on Knowledge Transfer to industry. The measure of Knowledge Transfer used is based on Woltmann and Alkearsig (2018), which traces identical content contained in publications and company websites using text mining applications.

Accordingly, we contribute to the academic debate in two ways: by providing insights into the changes in academic publishing behaviour over the years and Open Access implications for university-industry Knowledge Transfer. We also provide additional insights into companies' information identification strategies and knowledge adaptation. We make inferences about the usefulness of Open Access publishing in general and the policies that favour Open Access (for Science and Education, 2018; School of Electronics and Computer Science at the University of Southampton England, 2018). This helps to clarity whether restrictions on access to publications have consequences for industry R&D. It is important to have a better understanding of the true benefits of Open Access to provide a solid basis for an evaluation of related policies.

III.2 CONCEPTUAL FRAMEWORK

This section describes the principles underlying our understanding of the implications of Open Access. We need to clarify expectations and assumptions about the mechanisms underlying university-industry Knowledge Transfer and the implications of accessibility to this knowledge. Many empirical studies find Open Access publications to be beneficial. The advantages are measured mainly in terms of academic citations (Eysenbach, 2006;

Harnad and Brody, 2004) and have led to the assumption that the impact on the industry will be significant.

III.2.1 Background: Open Access Policies and Opinions

Academic publishing has evolved radically due to developments in Information Technology (IT) infrastructures, which have changed the entire publishing system. The costs of printing and distribution have decreased hugely, storage has become more feasible and dissemination of manuscripts has become (almost) instantaneous and free of costs (Laakso et al., 2011) all of which has sparked debate on the general rights related to accessing public research (Else, 2018). These developments have resulted in faster exchanges among scientists and led to accelerated knowledge creation and distribution (Tennant et al., 2016). Scholars from different disciplines and societal groups have been advocating for free availability of scientific research (Björk, 2004). For example, the university of Cambridge maintains that:

> 'The Open Research movement seeks to maximize the impact and benefits of research by prioritizing barrier-free access to research findings, data and methodologies. Open Research reflects a fundamental belief that the pursuit of knowledge benefits directly from collaboration, transparency, rapid dissemination and accessibility' [1].

The main arguments put forward are that traditional subscription based payment systems slow innovation by limiting scientific exchange, that the monopoly position of the biggest publishing houses has made them gate keepers to this knowledge and destroyed scientific diversity by fostering only mainstream thinking, and that high journal subscription prices have led to unfair competitive advantages for certain nations (Björk, 2004; Larivière et al., 2015). It has been pointed out also that the public, as main

1 https://www.cambridge.org/core/services/open-access-policies

funder of public research, should have the right to access research results (Armstrong, 2015; Arzberger et al., 2004). In this debate, concepts, such as Open Science and Open Access, as forms of scholarly communication, have become increasingly established and implemented (Jokić et al., 2018; Tennant et al., 2016).

Over time, the advantages and disadvantages for the academic community have been discussed and investigated by focusing mainly on the impact of Open Access on research quality and importance (citation counts) (Antelman, 2004). Relatively little has been published about the implications for society and industry; most of this literature consists of bibliometric analyses or normative opinion pieces based on qualitative data (McKiernan et al., 2016; Tang et al., 2017; X. Wang et al., 2015). Despite the lack of work on the impact of Open Access for society and industry, universities and governments have begun to act by implementing Open Access agendas and frameworks (Armstrong, 2015; Pinfield, 2015; Tennant et al., 2016). The extent of the global adoption of Open Access policies is evident in projects such as the Registry of Open Access Repository Mandates and Policies (School of Electronics and Computer Science at the University of Southampton England, 2018), which is an international registry for

> 'open access mandates and policies adopted by universities, research institutions and research founders that require or request their researchers to provide open access to their peer-reviewed research article output by depositing it in an open access repository' (School of Electronics and Computer Science at the University of Southampton England, 2018)

This means that policies are already in place, but are based on limited evidence. We need to take a step back and to address this knowledge gap to find evidence that either supports the positive assumptions about Open Access or shows that it is not living up to expectations. This study investigates this in the case of a Scandinavian university that held an external policy event which resulted in an increased focus on Open Access from 2012 (for

Independent Research et al., 2012). In 2018, a National Strategy for Open Access was announced in 2012 which put a stronger focus on Open Access and has changed the publishing reality in the national universities (for Science and Education, 2018).

We assume that policy framework changes (for Science and Education, 2018) have changed the nature of research dissemination in favour of Open Access publications. However, we need to verify whether Open Access publication is a sustainable trend and that there are no significant differences in terms of publication quantity and quality (highly relevant in academia). Hence, we hypothesize that:

Hypothesis 1a (H1a): *The ratio of freely available (Open Access) publications increases over time.*

Hypothesis 1b (H1b): *The quality of Open Access publications is comparable to quality of subscription based publications.*

III.2.2 Knowledge Transfer: Publishing Strategies

To understand the impacts it is important to understand the perspective of knowledge acquisition in industry. Research on knowledge acquisition and learning in industry is examined frequently with a company based view (Choo et al., 2001; Huber, 1991). Theories of organizational learning tend to be focused on companies learning from other companies via acquisitions or collaboration (Grant and Baden-Fuller, 2004). However, there are some approaches that consider the individual to be the main carrier of knowledge and see the organization as the user of this knowledge. According to Grant 'knowledge is viewed as residing within the individual, and the primary role of the organization is knowledge application rather than knowledge creation'. (Grant, 1996) [p. 109] Accordingly, individuals are perceived as the carriers, transmitters and receivers of specific information. Focusing on knowledge exchanged between individuals, groups and organizational units has given rise to the theoretical foundations for understanding Knowledge Transfer. This provides a

multitude of perspectives to aid in understanding why certain areas of knowledge see a more efficient transfer, and others not, across different organizational levels. A key aspect to explain the behaviour of the individual is information seeking (Krikelas, 1983).

It is important to differentiate between academic and industry researchers. Academic scientists, who spend a significant part of their working time monitoring academic output in their particular domains, are wary about the impact on their peers of citations and use of their research output (Cozzens, 1989; Thornley et al., 2015). In contrast, company researchers might just be seeking particular information rather than monitoring an entire field; however, the knowledge still flows and information seeking by company researchers follows similar patterns to those adopted by academic researchers (Ellis and Haugan, 1997).

Knowledge seeking is controlled by the knowledge receiver, who decides: a) where to search (systems and databases); b) which information to ignore; c) which information to absorb and utilize; d) which uses to display (Ho and F. Wang, 2015). Thus, the accessibility of information might play a crucial role for these researchers. Yet, it cannot be assumed that Open Access is the solution. It could be that free availability is perceived by company researchers as a sign of poor quality or unreliable research and might be exploited less (McCabe and Snyder, 2005). Nevertheless, not all companies have access to subscription based journals, which could hamper the information seeking by industry researchers (Armstrong, 2015). However, in our research setting this restriction would for most researchers only require time and effort to circumvent, if at all. It is important to consider the impacts of individual behaviour and search strategy at the meta level to understand the implications of freely available research information (Wilson, 2000).

While considering the entirety of information exchange, we need to consider receivers at the level of both the company and the individual researcher. To do this, we draw on aspects of communication theory, which is a mathematical representation

of the conditions and parameters affecting the transmission and processing of information. Communication theory is based on the work of the electrical engineer Claude Shannon (see Figure III.1). Some concepts have been adopted and are used in several fields including psychology and linguistics (Markowsky, 2017; Shannon, 1948).

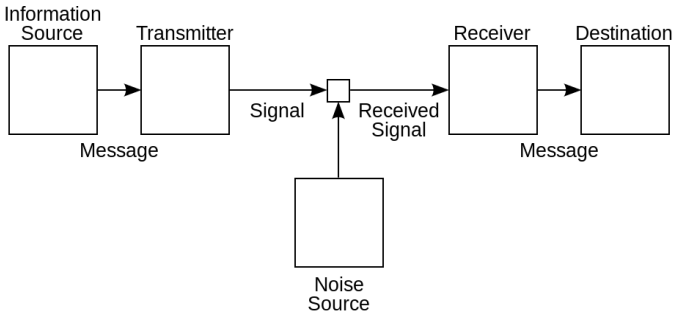


Figure III.1: Communication Model after Shannon 1948

This theory is particularly relevant because this study investigates two different modes of information transmission: Open Access and subscription based publications. There are three main message transmission types in computer networks: unicast, multicast and broadcast (see Figure III.2). Unicast transmission involves one node that sends the message to exactly one destination node (one-to-one transmission). Broadcast transmission refers to group communication, where the sender transmits a message addressed to multiple nodes without restriction, so called one-to-many communication. In this case, every node receives the message. Multicasting refers to messages sent only to a sub-group of potential receiving nodes and excludes some nodes in the network from receiving the message (Fairhurst, 2009; Nevase, 2016). Similar to a computer network, publishing methods map onto two types: broadcast and multicast.

Figure III.2: Unicast, Multicast, Broadcast after Viraj Nevase 2016

We need to consider these modes of communication and their senders when considering accessibility and information seeking behaviours, as they have a deciding role in the receiving process (Aurisicchio et al., 2010). The multicast model can be seen as the restricted subscription based publications, while an Open Access publication can be compared to broadcasting, as every node in the network can receive the message. However, we need to take account of the above mentioned behavioural aspects and the individual information seeking and adapting behaviour. This should add to our understanding of whether freely available information is received and used by more individuals, or whether restricted information has a greater influence and is valued more highly. Our objective is to understand whether the availability of research publications has implications for university-industry Knowledge Transfer; therefore, we need to examine potential differences in terms of Knowledge Transfer detection. There are two main aspects to this: first, research dissemination by universities and, second, adaptation behaviour of the potential receivers of the information, in our case, industry.

Using a combined model, we need to understand the constellation of university publications, which represent academic research information provided in different formats. Successful transfer can be measured only if the information is received and displayed. Therefore, we need to focus on the receiver. We are interested in whether private industry is more likely to use freely accessible or restricted information. However, we do not focus solely on the receiving aspect or the information seeking behaviour. We hypothesize that:

Hypothesis 2 (H2): *There is a higher ratio of Knowledge Transfer to industry from Open Access compared to subscription based publications.*

We need also to account for potential differences among different receivers (companies). Some might have formal or informal relations with universities - especially national universities - and this might influence Knowledge Transfer. This specific communication system of Knowledge Transfer might not independent from geographical and/or collaborative proximity of the two

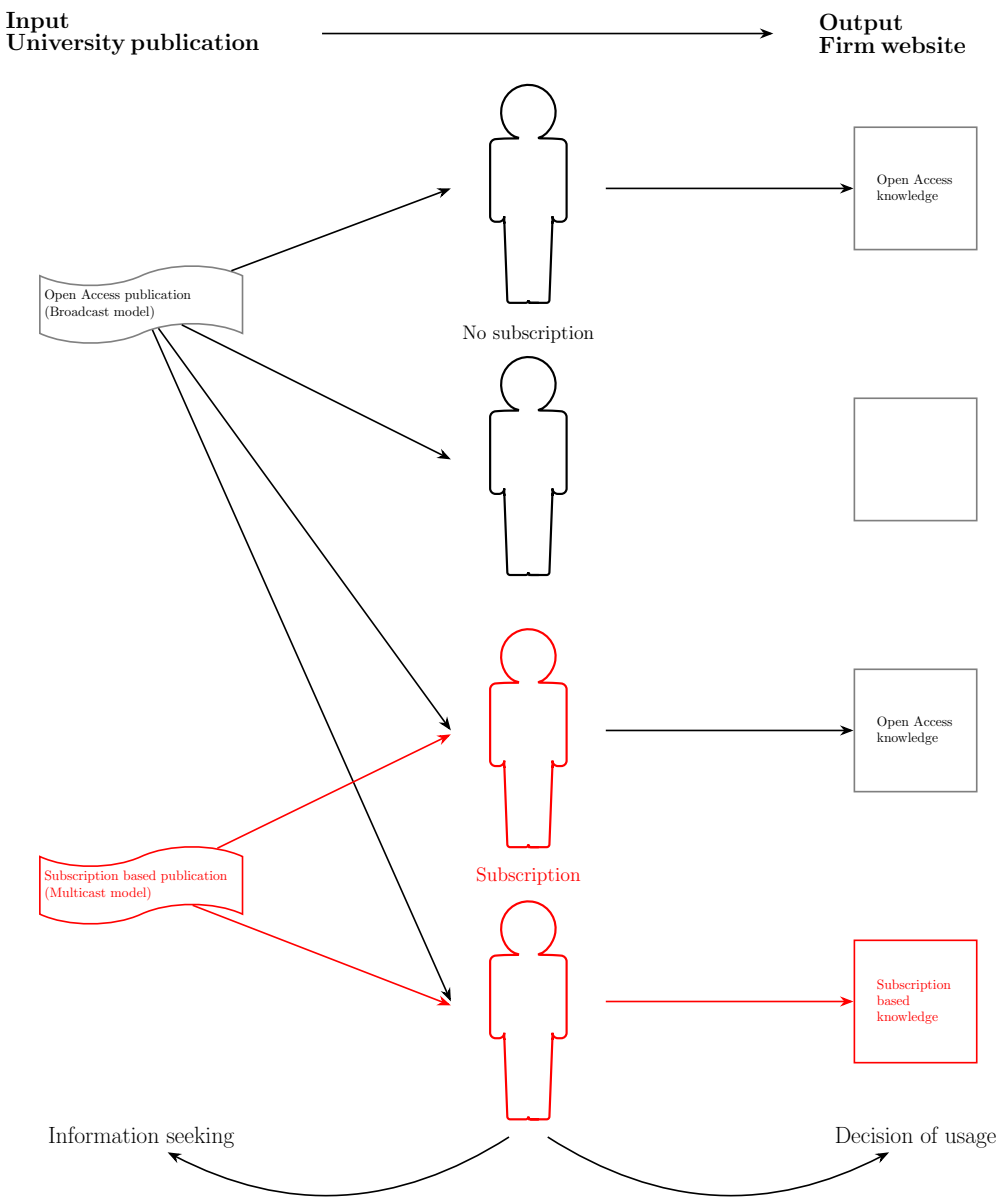


Figure III.3: Broadcast (Open Access) and Multicast (subscription) for university-industry Knowledge Transfer

components (Arundel and Geuna, 2004; Arundel, Geuna et al., 2001). We hypothesize that:

Hypothesis 3 (H3): *The rate of research Knowledge Transfer to companies is higher in the case of companies with a) formal ties to a university (contracts) and b) companies in the same national context (identified by a national registration number) as the university .*

III.3 RESEARCH STRATEGY

We address our research objective by testing the above hypotheses. This requires collection of data and statistical measures to confirm or reject them. We employ various strategies and statistical approaches. First, we want to show whether the data have the expected distribution and ratios (H1a); to test this, we chose to examine the entire publications sample of one technical university and the proportion of Open Access publications. Inevitably, publication age affects its transfer and the probability that it is an Open Access publication. Thus, we need to understand the changes overtime including, in particular, yearly distribution and particularities of respective research fields. We chose a data sample covering ten years (2005-2015), which includes the period when policy changes were announced, and happening. We test for equal proportions of Open Access or restricted publications, and apply a χ^2 test.

To verify H1b we use an indicator commonly used to measure publication quality and relevance: citation count. This shows how much a given publication is used within the scientific community (Meho, 2007; Moed, 2006). We consider it adequate to verify that the quality and relevance of Open Access versus subscription based publications are comparable. If we found significant variation, this inevitably could influence the final usage.

In line with our H2, we want to measure Knowledge Transfer from university to industry. This transfer should be related to specific publications, which allows us to identify relationships

between Open Access and transfer probability. To identify successful Knowledge Transfer, we apply the methods proposed in Woltmann (Woltmann and Alkærsig, 2018). We use company websites to identify company knowledge and compare it to the abstracts of university publications. By this means, we identify text pair matches, which later are validated manually. Testing H3 requires examination of an additional sample of websites (second degree), which represent companies that have no direct ties to the university, meaning they have no formal contract with the university. These are identified by website hyperlinks on the websites we identified based on formal ties (first degree sample). Finally, to validate our assumptions on more levels, we want to test whether combining the firstand second degree samples yields the same or comparable findings as the testing of H2. Our strategic approach does not investigate the transfer mechanisms, but does allow conclusions to be drawn about our conceptual model.

III.3.1 Definition of Open Access

Given the objectives of this study, we need to define Open Access. The 2002 Budapest Open Access Initiative (Budapest Open Access Initiative, 2002) defines Open Access as materials that are free to read, meaning without any subscription barriers and free to reuse (Budapest Open Access Initiative, 2002). The literature classifies Open Access into various types with different implications for the empirical research. The most common types are:

1. Green Open Access publications, which are published in paid access journals, but are self-archived in an Open Access archive and, hence, are available to everyone;

2. Gold Open Access, which includes articles that are made freely available by the relevant journal (usually paid for by the author) (Picarra et al., 2015);

3. Black Open Access, which refers to publications that are shared on illegal sites such as primarily Sci-Hub and LibGen (Tennant et al., 2016).

In our case we adapt the green Open Access definition and refer to Open Access in terms of articles available in the university repository.

III.3.2 Methods

To compare our two document types (university publication abstracts and company websites) we use an algebraic method that has proven efficient for these kinds of data, that is, Term-Frequency, Inverse Document Frequency (TFIDF). TFIDF is an established text mining numerical indexing method (Franceschini et al., 2016; Zhang et al., 2016) and can be used to extract keywords that rely on co-word occurrence. This approach has been shown to give most reliable results (Woltmann and Alkærsig, 2018). TFIDF indexing is used to determine the most frequent words (keywords) per document (maximum of 50) and then to compare each keyword list from the websites to each list of academic documents. We want, also, to identify differences in the companies investigated (see H3) using the TFIDF application.

Based on the findings from the previous study we set the same thresholds to limit the final verification work (see Section III.3.2). We built additional lists of keywords to allow a strategic exclusion of matches of not content related terms. These lists were mainly language fragments in French, German or Danish, country names or months. This resulted in a pre-processing phase which allowed us to limit our matches to those with content relevant keywords. For the comparisons, we used the Jaccard coefficient, which provides a number between 0 and 1 that defines the overlap in the items in the two sets. We set this threshold as a minimum 0.13. It is necessary to account for very short keyword lists that might have comparatively high Jaccard coefficients; we determined that a set of pairs with a Jaccard coefficient of less than 0.15

required more than seven words in common to meet the criteria (Woltmann and Alkærsig, 2018).

Human verification

Manual verification is needed to verify our outcomes for the hypotheses because there is no mechanical means available to confirm this the identical content. Human verification is currently the only way to ensure that the matches are chunks of common knowledge pieces. It is clear that the challenge to truly identify the content of a publication and a website and understand their overlap is yet to advanced for the contemporary computational methods. In particular with texts that are linguistically highly diverse. Hence, we use the computer generated text pair matches and verify them manually, which has been shown to require expert knowledge because of the complexity of the topics. Our manual checkers to assess the relatedness of the documents were individuals with higher education. To ensure a reliable outcome, we asked them to indicate their level of confidence in their assessment, which allowed for additional verification in the case of an insecure judgment. The verification involved three categories:

- First category: a verified match where the content of a publication and a website are identical;
- Second category: a match between the topics of the two documents;
- Third category: texts with nothing in common.

The size of the data set was problematic and did not allow generation of a labelled training/test set to facilitate computational assessment of potential errors. Also, even for experienced individuals, the verification was difficult; the distinction between true positives and false positives is very complex to allow reliance

on currently available computational methods. Also, findings are sparse and each finding is highly relevant to the output.

III.4 SAMPLE AND CASE

In this section, we discuss the decisions and outcomes of the data collection and samples. We focus explicitly on the data bases and their restrictions to provide an adequate overview of data quality and coverage. Section III.4.1 describes the university's (sender) publication data and Section III.4.2 focuses on the partner (receiver) data.

III.4.1 Publication data

We collected all the entries in the university's own publications data base (ORBIT) from 2005 to 2015. We retrieved meta-data including year, author(s) names, department (scientific field), publication type, etc. These meta-data provide the basis for any labelling and the later text mining procedures. We chose the time interval 2005-215 for several reasons:

- It covers the time when Open Access publishing became increasingly relevant (in policies, funding and for journals), hence, it provides the best insights into the impact of these changes. (Earlier than 2005 would risk a focus merely on subscription based publications which would defeat the purpose of the sample);

- 2015 needs to be the end date, since we cannot assume the same Knowledge Transfer outcomes for very recent publications. (The peak for citations generally occurs around 3 years after the publications date.), hence we assume one year as minimum for transfer as reasonable time frame for the Knowledge Transfer to the industry.

- The interval covers a period spanning from almost no Open Access publishing to a situation where more than half of the entries are Open Access, which provides a comparatively balanced sample.

- It is necessary to consider that the Open Access publications are mostly more recent than the subscription based publications, which, in terms of citations, could be a disadvantage, but in relation to our measure could be an advantage.

Open Access publications were extracted from the university's own data base with Open Access publications defined as all the papers available online and not restricted to any registered user access. To classify the entries into different research fields we used the department labels from the data base to assign them to a scientific field. We summarized the department labels into fields according to their scientific content. We categorized the 204 available department labels according to their overall scientific discipline. This was necessary to obtain labels that resembled the research content of the publications and not the university department, faculty and research group structure. This ruled out any effects of mergers, splits or renaming of research units. We obtained 20 distinct research fields and one collection of 'diverse' publications that could not reasonably be associated to any given field.

However, any interdisciplinary collaboration occurs only once in the data base and is in the data base assigned to the first author or in case of inter-university collaboration to the field in which the university employee is employed in. This classification is not an exact measure and a small number of articles might seem to be assigned arbitrarily. For the publication texts, we refer to the academic abstracts, which is a common approach in the field and has proven to give good results while at the same time maintaining feasible computation times. Overall, university publication numbers were fairly stable over the years 2005-2015 with a slight and steady increase in the number per year. This increase is likely due to incentives to publish more or to changes in the university structure. The distribution is comparatively

uniform, which suggest comparison across years is possible (see Figure III.6).

The distribution across research fields is more skewed. The total number of publications varies widely among research fields. This could be due to features such as age or size of the research unit, reliability of registration of publications, or publication intensity in the field in general. There are also substantial variations within single research fields over the years. Some fields show extreme publishing peaks and troughs, while others show continuous increases. This is not surprising given that constellations and focus areas change over the years and mergers or splits of university research units.

To infer publications data quality, we identified a sample of Open Access publications from the university in Scopus, a publications database regularly used in bibliometric studies. Scopus has been shown to be an adequate data-source in many previous empirical studies (Boyack, 2015; Kamdem et al., 2017). As our publication "quality" parameter we use overall citation counts, referring to the number of times a paper is cited. We compare citation counts for Open Access and subscription based publications. Recall that Open Access publications are generally more recent and have had less time to accumulate citations. However, recent studies show that publications citations tend to peak after 2-3 years (although this varies across research fields), which allows a general overview from the data within the chosen time frame.

III.4.2 First and Second Degree Partners

As described above, we want to compare texts of university research output to text documents containing company information. We focus on two different partner types that might affect Knowledge Transfer potential: first, companies with formal links to the university (contract) and a national registration, which we call first degree partners. Second, partners of first degree

partners that have no formal connection with the university and no national registration. The sampling of company related documents was performed by web crawling of company web pages during the years 2016 and 2017. The sample of companies was reduced by applying the following criteria:

1. The web pages have to be in English;

2. In the case of first degree companies, evidence of a connection to Denmark in the form of:

    - CVR number (Danish VTA number, meaning a company is registered in Denmark);

    - Or have the identifier ApS (Anpartsselskab) attached to the company name, indicating a Danish limited liability company.

First degree partners are companies with a formal contract with the university between 2006 and 2016. We identified a total 1,221 partners in a specific university contract data base; after applying the criteria, we obtained a sample of 541 first degree partner companies. The sample of web-pages and documents retrieved from the websites included 139,270 documents. For companies with at least five pages in English, each company has its own document archive; we obtained a final sample of 445 document collections.

We identified 41,289 potential second degree partners based on company websites; to achieve a more manageable and relevant sample, we retained the most relevant 28,000, from which we randomly sampled 1200 comparable company websites. To have included all 41,289 pages would have been excessive and would have made use of the more advanced text mining applications difficult or impossible. We assume that our sub-sample of 28,000 pages is representative of the total 41,289. For the second degree partners we used a sample of 1200 websites. To ensure a relevant sample from the random selection, we applied certain restrictions:

1. English content within the web-pages;

2. At least 5 web-pages with more than 60% of the content in English;

3. Website not included in the first degree sample.

Note, that company identifiers were not used to generate this sample. In other words, the websites might represent companies, but could belong to either group. We also did not apply geographical parameters. This approach is appropriate since we are interested in a random selection of any type of second degree partner. Also, removing the geographical restriction applied to the first degree sample might provide additional insights into the research knowledge diffusion.

III.5 RESULTS

In this section we discuss our statistical validation or rejection of our assumptions. The section is split into sub-sections according to our three hypotheses of the paper, which are followed by an examination of the combined data set. We also discuss our model choices.

III.5.1 Open Access Proportions and Quality

First, we collected all entries from the ORBIT database. For 2005-2015, we identified almost 24,000 Open Access publications, representing a third of the university's total publications. However, the distribution of Open Access publications over the years shows that they have increased (see Figure III.4). In particular, after 2009, this increase is very evidence and is a continuing trend. This general increase in Open Access publications is in line with

our expectations about the share of Open Access publications in the university's total publications.

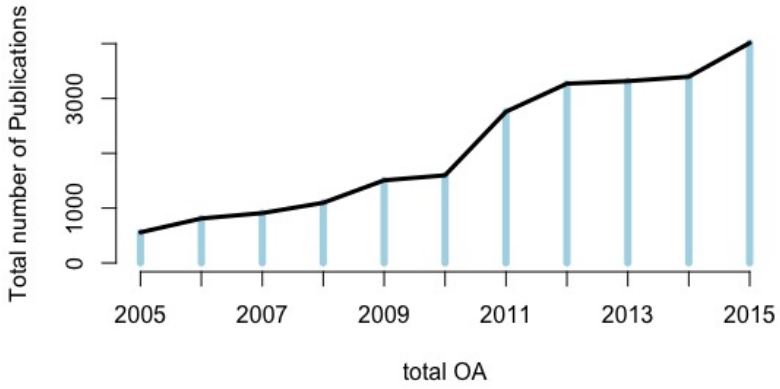


Figure III.4: Total OA publications of the university between 2005-2015

Although the university's total output is comparatively uniform, we observe a steady reduction in subscription based publications (see Figure III.5). This suggests that the university is producing an increasing number of freely available publications.

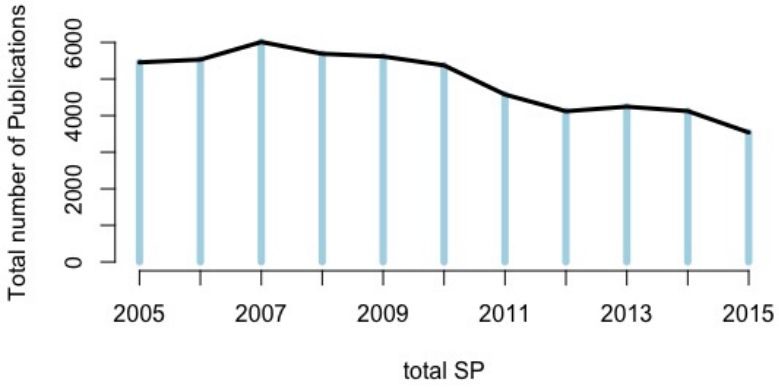


Figure III.5: Total subscription based publications publications of the university between 2005-2015

We applied a two-tailed test of the proportion of Open Access and subscription based papers, to assess the extend of the differences in these populations (see Table III.1). This establishes

comparability between the two groups. We need the overall probability to be not significantly different from 50% to allow comparability with other statistics. We assume the same distribution in both populations at the 95% confidence interval. Our results show that none of the fields has a number of Open Access publications that is beyond the population scope which means that they should be comparable in terms of Open Access and subscription based publications. However, this analysis applies only to the total number of papers and might vary when we consider yearly distributions.

Department	OPA	Total pub	SB	Pop. OA (p-value)
Aqua	787	4786	3999	0.164
BioSys	894	3535	2641	0.252
ChemBiochem	1027	4315	3288	0.238
Chemestry	389	2307	1918	0.168
CivilEng	995	3432	2437	0.289
ComputeMath	1928	5779	3851	0.333
Diverse	1949	8792	6843	0.221
ElectEng	1712	4271	2559	0.400
EngConSto	494	1492	998	0.331
EnviEng	1253	3824	2571	0.327
Food	1648	6194	4510	0.266
MAN	1886	4519	2633	0.417
MechEng	1223	4282	3059	0.285
MicroNano	900	3007	2107	0.299
NUC	198	418	220	0.473
PhotoEng	2058	5523	3465	0.372
Physics	679	1896	1217	0.358
Space Institute	780	2135	1355	0.365
Transport	470	1686	1216	0.278
Veterinary Institute	818	2583	1765	0.316
Wind	1156	1970	814	0.586
Total	23818	77920	54102	0.305

Table III.1: Journal Paper Distribution

If we look at the ratio between the two types of research publications we can see a rapid increase in Open Access publications.

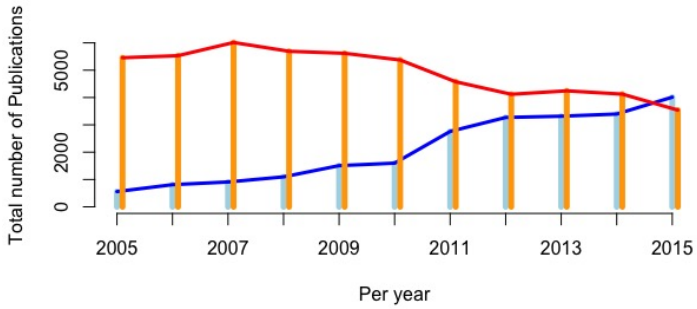


Figure III.6: Total publications of the university Research fields 2005-2015

From 2010, we observe a major shift towards more Open Access publications (see Figure III.6), with 2015 the first year when the university produced more open access than subscription based outputs. .

Publication year	Total publications	total OA	Ratio
2005	6046	587	0.10
2006	6368	823	0.13
2007	6957	916	0.13
2008	6822	1108	0.16
2009	7183	1533	0.21
2010	7050	1614	0.23
2011	7409	2781	0.38
2012	7491	3305	0.44
2013	7665	3385	0.44
2014	7647	3464	0.45
2015	7789	4197	0.54

Table III.2: Paper Distribution over years

The 20 research fields show changes in the distribution of Open Access and subscription based publications. However, we can see that individual research fields are not aligned to the overall picture and there is some variation both between and within fields. All fields start out in 2005 with a very low share of or no Open Access publications. This increases during the next years, but the extent of the increase varies across fields. The

distribution over time shows that the two types of publications are comparable along our time line, but that 2005-2007 were close to the threshold that needs to be considered when interpreting the results (see III.2). Some research fields show significant peaks and troughs, which makes their trend less consistent, but it should be noted that, in some fields (with small numbers of publications), the ratio might be affected by the addition of only a small number of publications (e.g. Wind energy and Energy Conversion and Storage). Most Open Access publications are generated by the most active publishing fields, however, around half of them display the same distribution over the years as the general trend. In 2015, more than half of the publications in eight fields are freely available. In three cases, more than half of their publications were freely available prior to 2015. (see Figures III.7 and III.8).

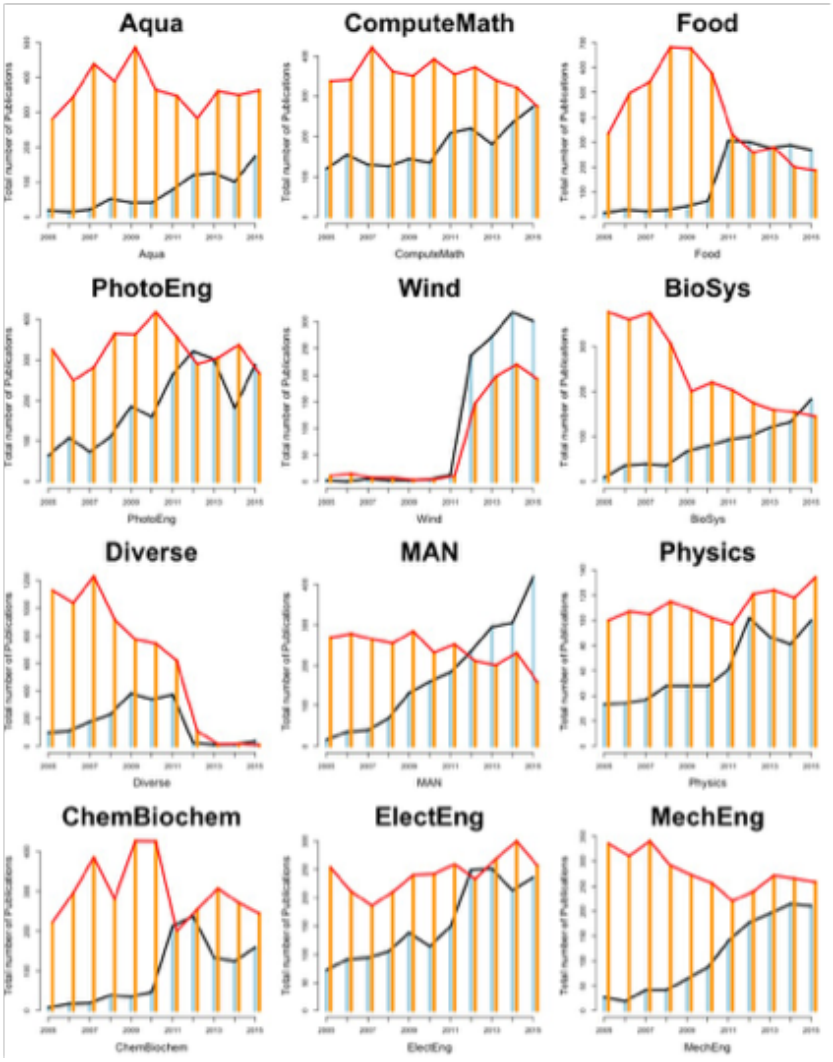


Figure III.7: Open Access vs subscription based publications of the university Research fields 2005-2015

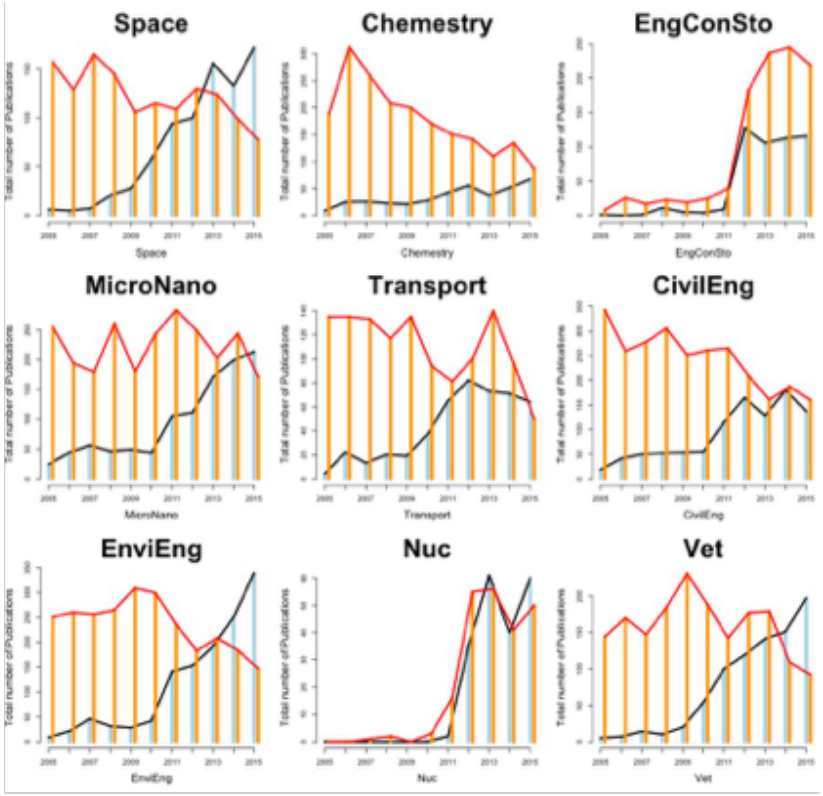


Figure III.8: Open Access vs subscription based publications of the university Research fields 2005-2015

We also checked the proportions of years and fields. The results show that comparing individual departments for each year is not feasible, especially for years 2005 to 2008. In some years, Space research, Nuclear technologies and Wind energy have zero Open Access publications and in most of the other fields, the numbers are very low prior to 2008. This unevenness in the distribution shows that comparison of fields including years is not feasible. Hence, in the following analysis, we include years and the fields as separate variables. Overall, our results confirm H1a with some exceptions for a few particular academic fields. Figure 9 depicts numbers of citations (in-degree distribution) to Open Access and subscription based publications. Although only 25% of the papers are Open Access, they appear to as visible and to be quoted as often as subscription based journal publications. However, only some 30%-40% of both Open Access and subscription based publications are registered in Scopus

(Woltmann et al., 2018). Hence, we can assume that in both cases, a similar share of the sample set is represented and our results for quality hold and do not favour either sample in particular. Comparing Open Access to subscription based entries shows that, in terms of in-degree citations (references to this paper), Open Access and subscription based are of comparable scientific importance (see Figure III.9).

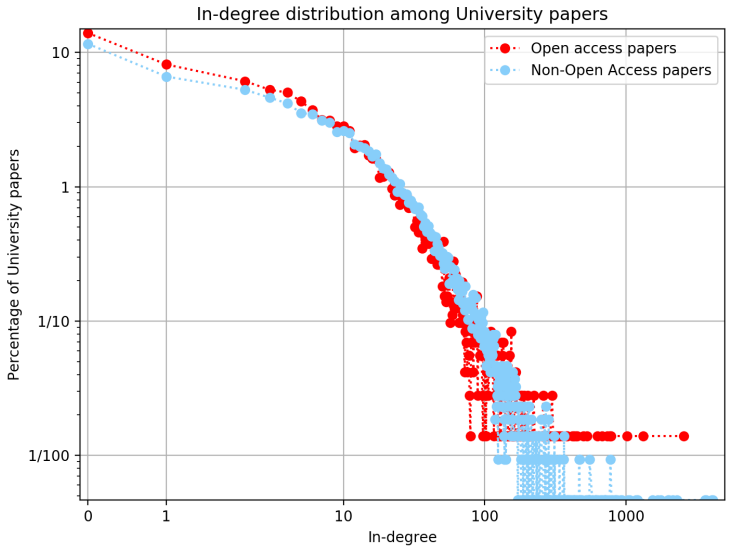


Figure III.9: Open Access vs subscription based publications of the university Research fields 2005-2015

III.5.2 Knowledge Transfer First Degree Partners

We applied text mining methods to the academic publications and company websites to enable comparison between the two types. First, we used the TFIDF indexing method for the first degree company websites, which resulted in 91 potential hits and a high number of research fields (see Table III.5.2). The overall number of verified matches shows that Open Access has more than than 50% and dominates the publications present Knowledge Transfer. This is especially relevant since the most relevant fields do not correspond to the fields with the highest Open

Access ratios. However, they are the most publication intensive fields. Testing the relevance of the field in relation to Knowledge Transfer, using occurrence of fields within the positive matches based on a χ^2 test, we find that field has no significant influence on Knowledge Transfer $p = 0.8$ and $\chi^2 score = 5.4112$). However, given our small sample size and the large number of potential fields this might change for a larger sample. It should be noted that the first degree partner websites were collected at one point in time in 2016. Table III.5.2 shows the distribution of Knowledge Transfer matches, identified using both computational methods and manual verification. It shows that the overlap in topics and the computational identified matches are very high, but the verification is still needed for the first category.

Method	Field (fields)	pot. Match	verified M.	OA of Verifed	topic (fields)	OA of topic
	Total (nr. 15)	91	23	12 (52%)	46 (nr.13)	9 (22%)
TFIDF	**Aqua**	1	1	-	1	-
	ChemBiochem	3	1	-	2	1
	ComputeMath	6	-	-	2	-
	Diverse	4	-	-	3	-
	ElectEng	9	1	1	6	-
	EngConSto	2	-	-	1	1
	EnviEng	2	1	1	1	1
	Food	14	2	1	8	1
	MAN	9	2	1	4	2
	MechEng	6	2	1	2	1
	MicroNano	3	1	-	2	-
	PhotoEng	17	9	5	7	1
	Space	7	3	2	4	1
	Vet	3	-	-	3	-
	Wind	4	-	-	-	-

To understand whether Open Access publications contribute to Knowledge Transfer, we run logistic regression models. In addition to the pure ratio distribution, we looked at the spread across different research fields. We tested different potential logit models to explain our Knowledge Transfer detection. We included as independent variables research field, Open Access (as binary true or false variable) and the publication age (publication year). We chose research field and age as relevant variables based on the relevance identified in prior empirical studies of research areas and research fields for Knowledge Transfer. We identified the best models, using the Akaike Information Criterion (AIC), which is an estimator of the relative quality of statistical models. Our approach considered all variables and potential integration effects; we limited the output to the five best models.

Our findings suggest that Knowledge Transfer depends on publication age (Model 1) and publication age and Open Access as independent variables (Model 2). However, Open Access is not significant. Models 3 shows an interaction effect of Open Access and publication age with research field as the independent variable, which is significant. Model 4 shows that there is a potential interaction effect between research field and publication year (see Table III.3). We consider all four models relevant since model performance differed only marginally. In our case, the AIC shows less than a one point difference, meaning that the explanatory power of the models is comparable.

It seems that Open Access is not the main driver of Knowledge Transfer. However, based on the distribution of Open Access publications in our data set, we included the variable publication year in Model 2. The outcomes required several adaptations to investigate potential relationships in depth, since publication age might affect the measurement of Knowledge Transfer. We want to ensure that the observed effect is not due to the younger age of the Open Access publications. Hence, we included years (2005-2015) as an additional independent variable.

For this sample no relevant effects apart from year are identified. However, given the increasing relevance of Open Access

Model 1	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			77515	419.64	
year	1	7.98	77514	411.66	0.0047***
Model 2	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			77515	419.64	
year	1	7.98	77514	411.66	0.0047***
OpenAccess	1	1.91	77513	409.75	0.1672
Model 3	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			77515	419.64	
OpenAccess*year	2	9.88	77513	409.76	0.0071***
Model 4	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			77515	419.64	
department2*year	21	47.52	77494	372.12	0.0008***

Table III.3: Model tests for Model 1-4 (first degree partners): Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

after 2010, due to policy and funding changes, we decided to use a different time frame for the statistical analysis and to run the logistic regression on a different data set that included entries only from 2010 onward. This data set generated the Models 5 and 6 (see Table III.4) which also include Open Access and research field as independent variables. Since Open Access publications are more frequent from 2011, we decided to investigate whether using the frame 2011-2015 changed our results. This subset of the original sample contains 34,326 entries, among which 16753 are Open Access publications (44%), and more than 80% of the potential matches. The best models are again similar in terms of the AIC. However, there is no single variable that seems relevant for this subset. In both models (Models 5 and 6) we can see that the most important variable is the interaction effect between year and Open Access.

Given that our sample includes all types of codified outcome produced by a university employee, we decided to check for the most relevant and dominant research type: journal publication (Gisvold, 1999). This choice is reasonable since journal publications is are the most important means of knowledge diffusion among scientists and universities. Journal publications are used

Model 5	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			37355	326.18	
year	4	4.43	37351	321.75	0.3515
OpenAccess	1	1.33	37350	320.42	0.2481
year*OpenAccess	4	13.44	37346	306.97	0.0093**
Model 6	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			37355	326.18	
OpenAccess	1	1.30	37354	324.88	0.2542
OpenAccess*year	8	17.90	37346	306.97	0.0220**

Table III.4: Model tests for Model 5 and 6 (first degree partners): Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

to evaluate research quality, novelty and importance. Hence, we assume imposing a further limitation on this parameter would benefit the analysis. The remaining subset of data covers some 34,326 (44%) of the original 77,516 entries. This sample contains 7,601 (33%) of the 23,233 Open Access publications and more than half (61%) of the verified matches. In this sub-sample, research field seems to have no additional value for the input model since even the null model performs better than the full model including department. Overall, Open Access is clearly a significant parameter (see III.5).

Model 7	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			34325	246.52	
OpenAccess	1	8.01	34324	238.51	0.0047***
Model 8	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			34325	246.52	
year	10	17.89	34315	228.63	0.0568*
OpenAccess	1	4.81	34314	223.82	0.0283**

Table III.5: Model tests for Model 7 and 8 (first degree partners): Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

We can see that it is only under certain conditions that Open Access is related to the Knowledge Transfer identified. This provides partial support for our hypotheses, in particular, in relation to Open Access versus subscription based journal publications. However, it shows that the p values are not completely

satisfactory and additional variables might have significant influence on the models. Hence, we cannot fully rule out that we are observing an artifact of this particular data set.

III.5.3 Second Degree Partners

After additional cleaning, the second degree sample includes 1,058 websites. We again compared each keyword list (derived from TFIDF) from every page of these websites with each keyword list of an academic abstract. Overall, we identified 118 potential matches based on our thresholds. The distribution of potential matches is comparatively skewed towards certain research fields (see Table III.5.3). From the full matches and content related entries (together 65), only 27 are Open Access, that is 41%. We see also comparatively high rates of false positives and very few verified matches.

Method	Field (fields)	pot. Match	verified M.	OA of Verified	3rd cat. (fields)	OA of 3rd
		Total 118	8 (7%)	3 (38%)	62 (52%)	24 (39%)
TFIDF (2nd degree)	Aqua	2	-	-		
	BioSys	4	-	-		
	ChemBiochem	1	-	-		
	Chemestry	1	-	-		
	ComputeMath	28	-	-		
	Diverse	20	1	1		
	ElectEng	10	-	-		
	EnviEng	1	1	-		
	Food	8	-	-		
	MAN	13	1	-		
	MechEng	8	1	-		
	MicroNano	6	3	1		
	PhotoEng	8	-	-		
	Transport	4	1	1		
	Vet	1	-	-		
	Wind	3	-	-		

Model 9	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			37355	326.18	
year	4	4.43	37351	321.75	0.3515
OpenAccess	1	1.33	37350	320.42	0.2481
year*OpenAccess	4	13.44	37346	306.97	0.0093
Model 10	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			37355	326.18	
OpenAccess	1	1.30	37354	324.88	0.2542
OpenAccess*year	8	17.90	37346	306.97	0.0220
Model 11	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			37355	326.18	

Table III.6: Best models (9-11) for the subset of 1st and second degree with all entries: Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

Comparing the 118 potential matches among the second degree partners to the 91 potential matches in the first sample that the overlap is less high within the second degree partners. This confirms that relationship and proximity make a difference for the identification of potential matches ($p = 1.982e^{-06}$ and $\chi^2 score = 22.612$). It shows that first degree partners are closer to academic outputs than second degree partners. Either the measures and thresholds perform less well in the second sample or, as seems more likely, the content is less relevant in the case of no formal partnership.

To confirm and add to our previous findings, we test the same models for the concatenated data set of first and second degree identified matches. We tested for the best logit models for this subset including all the independent variables (Open Access, research field, publication age). Our tests show that the model that includes Open Access and year is the best model for this subset to predict Knowledge Transfer (see Table III.6. However, if we include only Open Access, the AIC is very similar (AIC Model 9: 347.57 and AIC: 347.63 Model 10) and highlights the clear importance of Open Access and publication age.

For the second data set, which contains all entries from 2010 onwards, we tested for the most relevant models. We find the

first indication that research field matter for Knowledge Transfer (see Table III.7). However, we see no evidence that Open Access plays any significant role.

Model 12	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			37355	400.79	
department2	20	42.18	37335	358.61	0.0026 ***
Model 13	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			37355	400.79	
OpenAccess	1	0.84	37354	399.96	0.3601
department2	20	41.76	37334	358.19	0.0030***

Table III.7: Best models for the subset of 1st and second degree with only entries from 2010: Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

The final set of tests are run on the sample containing only journal publications (see Table III.8). In line with the findings for the previous sample of first degree partners, we find that Open Access is a relevant variable for Knowledge Transfer. This confirms our first finding for the first degree sample.

Model 14	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			34325	352.75	
year	10	24.01	34315	328.74	0.0076***
OpenAccess	1	5.17	34314	323.57	0.0230**
Model 15	**Df**	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
NULL			34325	352.75	
OpenAccess	1	9.12	34324	343.63	0.0025***

Table III.8: Best models for the subset of firstand second degree with only Journal entries: Signif. codes: 0.01 '***' 0.05 '**' 0.1 '* ' 1

Our analysis of the results of the first and the second degree samples revealed some interesting findings. When we examined the 'content' overlap in the first and second degree samples we found, first, the first degree and second degree partner research fields show strong overlaps (see Table III.5.2). This applies less to the second degree sample (see Table III.5.3). Second, we find

that Open Access is not significantly more related to industry in either case.

Third, we found an overlap for academic papers and potential matches between the firstand the second degree partners. There are six overlaps in potential document matches, that is, among academic documents that occur in the respective 91 and 118 firstand second degree partners. These overlaps are in the fields of:

- Compute Mathematics (2 papers);
- Management;
- Food Science;
- Wind Energy;
- Veterinarian Institute

These six papers gave for the second degree partners 11 potential hits. In particular, none were classified as verified Knowledge Transfer matches, but they indicate a high similarity of topic to the industry websites.

III.6 DISCUSSION

Our hypotheses are broadly supported: we found an increased orientation towards Open Access publications in university output and found also that, in certain cases, Open Access publications are more likely to result in successful university-industry Knowledge Transfer. We found general support for H1a and H1b. The findings for H1a show a clear increase in the ratio of Open Access publications, which is in line with our expectations, the academic and empirical literature and public policies (Piwowar et al., 2018). It is interesting that, in the last year of

our data (2015), Open Access overtook traditional subscription based publications as the predominant method of knowledge dissemination for the university. In 2015, Open Access publications represented more than half of the total publications for that year, which is above the estimated average of 45% of the global research publications (Piwowar et al., 2018). The time frame in our study is coherent with the time frames used in the literature. Our findings should convince policy makers to pay more attention to Open Access publication (Tennant et al., 2016). In addition, our findings confirm H1b, finding that the quality of Open Access publications does not differ significantly from their subscription based counterparts.

The results for both H1a and H1b lead to three main insights. First, the importance of Open Access is increasing based on its overall share, further highlighting the importance of understanding this phenomenon. While part of this effect is certainly caused by policy changes, it highlights a general shift away from the subscription based publications by university researchers. This contributes to the potential knowledge dissemination of university science through broadcasting, potentially making this knowledge more accessible to a wider audience.

Second, shares of Open Access could become so high that comparative studies will become impossible because of their skewed distribution. Should the trend we studied continue, the vast majority of publications in the future will be Open Access. This will be directly opposed to the dominance of subscription based publications in the past decades, making the current window of observation potentially the only frame in which the effects of Open Access publications can be compared directly to subscription based publications in this manner.

Third, the quality of Open Access and subscription based papers is shown to be comparable. It could be argued that researchers publish only their very best or lowest quality results as Open Access, with an underlying assumption that researchers are choosing to let the most promising research reach as wide an audience as possible, or in the opposite end of the spectrum,

using Open Access journals as a quicker method of publication compared to subscription based outlets (McCabe and Snyder, 2005). However, this study confirms that is not the case. As evidences by the findings supporting H1b, we find a wide variety of quality in both Open Access and subscription based publications, highlighting that both types of outlets is the target of research across the spectrum of quality.

Our findings provide support for H2 only under certain conditions. We hypothesised that Open Access publishing would be more likely to lead to Knowledge Transfer. This hypothesis is partially confirmed under certain conditions, however is not verifiable in our full sample. When examining the full sample of all types of documents, we did not find evidence that Open Access publishing would lead to increased Knowledge Transfer. However, when examining journal publications only, Open Access is found to be a relevant factor in Knowledge Transfer. Journal publications are perceived as the most relevant indicator of university output, and the primary research output of the university compared to other document types such as media articles and books. While we cannot universally confirm H2, the fact that this effect is present exactly for journal publications indicates that using the broadcast mechanism of Open Access to disseminate research results has a stronger effect than the multicast mechanism inherent to subscription based outlets. In this case, broadcast allows knowledge to have a higher visibility to information seekers, increasing the likelihood that the knowledge is received and utilized. In addition, publication date age and Open Access show interaction effects; further distinguishing between these two effects would be relevant.

Our findings for the relations among topics provide no clear evidence that Open Access publications have more content of interest to industry than subscription based publications. This might mean that Open Access is more relevant for industry; it might be based on strategic selection by researchers or might be a structural artifact based on research field and industry area. However, this aspect requires further investigation, using the entire sample and not just potential matches.

In the case of H_3 we found evidence that formally related companies located in the same region, show a significantly higher level of Knowledge Transfer. Although our identification method have not captured all relevant Knowledge Transfer activities, the strong significance of this finding is generalizable, highlighting the importance of (geographical) proximity or formal connections for information seekers. However, we found no differences between Open Access and subscription based publications. The logistic regression models were based on the firstand second degree samples combined; the second degree sample seemed to enhance the first observations.

The yearly fluctuations in different research fields and in the distribution of Open Access make it hard to exclude the other variables. In addition, the website collection time frame might be having some effect. Since the sample is very sparse and small, it is hard to exclude that some of the findings may be artifacts of these conditions. We cannot exclude that Open Access is only one among several characteristics such as research quality or topic of the publications. The implications of these findings are threefold; For policy makers, a continued push for Open Access broadcast of university research should be continued.

As highlighted by this study, Open Access journal publications have a higher likelihood of transferring knowledge to industry, and as such, efforts to push for an increased ratio of Open Access should revolve around these. Companies can find applicable knowledge both in Open Access and subscription based publications, as evidenced by our study. As such, to enable the most efficient information seeking, industry researchers should continue to search for information in both types of publications. However, companies without access to subscription based publications, can still find relevant information by searching only Open Access outlets as evidenced by our study. University researchers, while in many cases pushed to publish increasingly in Open Access outlets, can continue the current trend without limiting the potential impact of their research to industry. However many other factors determine whether Knowledge Transfer is success-

ful, and as such, Open Access publication should be perceived as an aid in knowledge dissemination, rather than a determinant.

III.7 CONCLUSION

Overall, our findings indicate that both Open Access and publication age affect university-industry Knowledge Transfer. However, given the high relevance of Open Access in contemporary research policies, funding parameters and university strategies, we expected greater significance of Open Access. Our empirical investigation should be extended with additional empirical measures in future research. The assumptions that accessibility is related to increased use are confirmed in certain circumstances. The implications for the private sector might vary from current assumptions. This calls for theoretical adjustments to focus and perceptions and additional explanations and theoretical concepts. Future research should disentangle the influence of publication novelty and Open Access. Our positive assumptions about the impact of Open Access could not be confirmed. Given the current trends towards Open Access, it should be acknowledged that measuring the true effects of Open Access is problematic due to the externalities involved.

REFERENCES – MANUSCRIPT III

Agrawal, Ajay (2001). 'University-to-industry knowledge transfer: Literature review and unanswered questions'. In: *International Journal of management reviews* 3.4, pp. 285–302.

Antelman, Kristin (2004). 'Do open-access articles have a greater research impact?' In: *College & research libraries* 65.5, pp. 372–382.

Armstrong, Mark (2015). 'Opening access to research'. In: *The Economic Journal* 125.586, F1–F30.

Arundel, Anthony and Aldo Geuna (2004). 'Proximity and the use of public science by innovative European firms'. In: *Economics of Innovation and new Technology* 13.6, pp. 559–580.

Arundel, Anthony, Aldo Geuna et al. (2001). *Does proximity matter for knowledge transfer from public institutes and universities to firms?* Tech. rep. SPRU-Science and Technology Policy Research, University of Sussex.

Arzberger, Peter, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir and Paul Wouters (2004). 'Promoting access to public research data for scientific, economic, and social development'. In: *Data Science Journal* 3, pp. 135–152.

Aurisicchio, Marco, Rob Bracewell and Ken Wallace (2010). 'Understanding how the information requests of aerospace engineering designers influence information-seeking behaviour'. In: *Journal of Engineering Design* 21.6, pp. 707–730.

Bishop, Kate, Pablo D'Este and Andy Neely (2011). 'Gaining from interactions with universities: Multiple methods for nurturing absorptive capacity'. In: *Research Policy* 40.1, pp. 30–40.

Björk, Bo-Christer (2004). 'Open access to scientific publications–an analysis of the barriers to change?' In:

Boyack, Kevin W (2015). 'Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database.' In: *ISSI*.

Budapest Open Access Initiative (2002). *Budapest Open Access Initiative*. URL: https://www.budapestopenaccessinitiative.org/read.

Choo, Chun Wei et al. (2001). 'Environmental scanning as information seeking and organizational learning'. In: *Information research* 7.1, pp. 7–1.

Cozzens, Susan (1989). 'What do citations count? The rhetoric-first model'. In: *Scientometrics* 15.5-6, pp. 437–447.

D'Este, Pablo and Parimal Patel (2007). 'University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry?' In: *Research policy* 36.9, pp. 1295–1313.

Ellis, David and Merete Haugan (1997). 'Modelling the information seeking patterns of engineers and research scientists in

an industrial environment'. In: *Journal of documentation* 53.4, pp. 384–403.

Else, Holly (2018). 'Radical open-access plan could spell end to journal subscriptions.' In: *Nature*.

Eysenbach, Gunther (2006). 'Citation advantage of open access articles'. In: *PLoS biology* 4.5, e157.

Fairhurst, Gorry (2009). 'Unicast, Broadcast, and Multicast'. In:

For Independent Research, The Danish Council, Danish Council for Strategic Research, Danish National Research Foundations, Danish Advanced Technology Foundation, the Danish Council for Technology and Innovation (2012). 'Open Access policy for public-sector research councils and foundations'. In:

For Science, Danish Agency and Higher Education (2018). 'Denmark's National Strategy for Open Access'. In:

Franceschini, Simone, Lourenço GD Faria and Roman Jurowetzki (2016). 'Unveiling scientific communities about sustainability and innovation. A bibliometric journey around sustainable terms'. In: *Journal of Cleaner Production* 127, pp. 72–83.

Geuna, Aldo and Alessandro Muscio (2009). 'The governance of university knowledge transfer: A critical review of the literature'. In: *Minerva* 47.1, pp. 93–114.

Gisvold, Sven-Erik (1999). 'Citation analysis and journal impact factors–is the tail wagging the dog?' In: *Acta anaesthesiologica scandinavica* 43.10, pp. 971–973.

Grant, Robert M (1996). 'Toward a knowledge-based theory of the firm'. In: *Strategic management journal* 17.S2, pp. 109–122.

Grant, Robert M and Charles Baden-Fuller (2004). 'A knowledge accessing theory of strategic alliances'. In: *Journal of management studies* 41.1, pp. 61–84.

Hajjem, Chawki, Stevan Harnad and Yves Gingras (2006). 'Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact'. In: *arXiv preprint cs/0606079*.

Harnad, Stevan and Tim Brody (2004). 'Comparing the impact of open access (OA) vs. non-OA articles in the same journals'. In: *D-lib Magazine* 10.6.

Ho, Mia Hsiao-Wen and Fatima Wang (2015). 'Unpacking knowledge transfer and learning paradoxes in international strategic

alliances: Contextual differences matter'. In: *International Business Review* 24.2, pp. 287–297.

Huber, George P (1991). 'Organizational learning: The contributing processes and the literatures'. In: *Organization science* 2.1, pp. 88–115.

Jokić, Maja, Andrea Mervar and Stjepan Mateljan (2018). 'Scientific potential of European fully open access journals'. In: *Scientometrics* 114.3, pp. 1373–1394.

Kamdem, Jean P, Kleber R Fidelis, Ricardo G S Nunes, Isaac F Araujo, Olusola O Elekofehinti, Francisco A B da Cunha, Irwin R A de Menezes, Allysson P Pinheiro, Antonia E Duarte and Luiz M Barros (2017). 'Comparative research performance of top universities from the northeastern Brazil on three pharmacological disciplines as seen in scopus database'. In: *Journal of Taibah University Medical Sciences* 12.6, pp. 483–491.

Kochenkova, Anna, Rosa Grimaldi and Federico Munari (2016). 'Public policy measures in support of knowledge transfer activities: a review of academic literature'. In: *The Journal of Technology Transfer* 41.3, pp. 407–429.

Krikelas, James (1983). 'Information-seeking behavior: Patterns and concepts.' In: *Drexel library quarterly* 19.2, pp. 5–20.

Laakso, Mikael, Patrik Welling, Helena Bukvova, Linus Nyman, Bo-Christer Björk and Turid Hedlund (2011). 'The development of open access journal publishing from 1993 to 2009'. In: *PloS one* 6.6, e20961.

Larivière, Vincent, Stefanie Haustein and Philippe Mongeon (2015). 'The oligopoly of academic publishers in the digital era'. In: *PloS one* 10.6, e0127502.

Markowsky, George (2017). 'Information Theory'. In:

McCabe, Mark J and Christopher M Snyder (2005). 'Open access and academic journal quality'. In: *American Economic Review* 95.2, pp. 453–459.

McKiernan, Erin C, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg et al. (2016). 'How open science helps researchers succeed'. In: *Elife* 5.

Meho, Lokman I (2007). 'The rise and rise of citation analysis'. In: *Physics World* 20.1, p. 32.

Moed, Henk F (2006). *Citation analysis in research evaluation*. Vol. 9. Springer Science & Business Media.

Nevase, Viraj (2016). 'A Practical Guide to Differentiate Unicast, Broadcast & Multicast'. In: URL: https://www.esds.co.in/blog/difference-between-unicast-broadcast-and-multicast/.

Picarra, Mafalda, EKT Victoria Tsoukala and Alma Swan (2015). 'Open Access to scientific information: facilitating knowledge transfer and technological innovation from the academic to the private sector'. In: *PASTEUR4OA Briefing Paper*.

Pinfield, Stephen (2015). 'Making open access work: The 'state-of-the-art' in providing open access to scholarly literature'. In: *Online Information Review* 39.5, pp. 604–636.

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West and Stefanie Haustein (2018). 'The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles'. In: *PeerJ* 6, e4375.

School of Electronics and Computer Science at the University of Southampton England (2018). *Registry of Open Access Repository Mandates and Policies*. URL: https://roarmap.eprints.org/.

Shannon, Claude Elwood (1948). 'A mathematical theory of communication'. In: *Bell system technical journal* 27.3, pp. 379–423.

Tang, Min, James Bever and Fei-Hai Yu (2017). 'Open access increases citations of papers in ecology'. In: *Ecosphere* 8.7.

Tennant, Jonathan P, François Waldner, Damien C Jacques, Paola Masuzzo, Lauren B Collister and Chris HJ Hartgerink (2016). 'The academic, economic and societal impacts of Open Access: an evidence-based review'. In: *F1000Research* 5.

Thornley, Clare, Anthony Watkinson, David Nicholas, Rachel Volentine, Hamid R Jamali, Eti Herman, Suzie Allard, Kenneth J Levine and Carol Tenopir (2015). 'The role of trust and authority in the citation behaviour of researchers.' In: *Information Research* 20.3.

Tijssen, Robert (2006). 'Universities and industrially relevant science: Towards measurement models and indicators of entrepreneurial orientation'. In: *Research Policy* 35.10, pp. 1569–1585.

Wang, Xianwen, Chen Liu, Wenli Mao and Zhichao Fang (2015). 'The open access advantage considering citation, article usage and social media attention'. In: *Scientometrics* 103.2, pp. 555–564.

Wilson, Thomas D (2000). 'Human information behavior'. In: *Informing science* 3.2, pp. 49–56.

Woltmann, Sabrina, Lars Alkærig and Carina Lomberg (2018). 'Open Access' influence on University-Industry Knowledge Transfer'. In: *In PhD Thesis*.

Woltmann, Sabrina and Lars Alkærsig (2018). 'Tracing university–industry knowledge transfer through a text mining approach'. In: *Scientometrics*. URL: https://doi.org/10.1007/s11192-018-2849-9.

Zhang, Yi, Lining Shang, Lu Huang, Alan L Porter, Guangquan Zhang, Jie Lu and Donghua Zhu (2016). 'A hybrid similarity measure method for patent portfolio analysis'. In: *Journal of Informetrics* 10.4, pp. 1108–1130.

RESULTS NOT PRESENTED IN THE MANUSCRIPTS

This separate section describes in brief some applications of computational methods on the data samples of Manuscript II and Manuscript III. Only the preliminary results and considerations are described and explained. I add this section to ensure that other approaches in future can take these insights into consideration.

Principal Component Analysis (PCA)

To investigate and understand the data and its properties, I applied a principal component analysis (PCA) and a SPCA to the text data of the websites. The results were interesting and showed noticeable groupings.

The PCA was mainly applied to reduce the dimensions of the sample and to examine if there are interesting trends in the texts to be identified. The PCA was applied on the document as well as the term dimension of the matrix. To gain additional insights, I tried using different weighting schemes additionally to the TFIDF calculation. (see Chapter 6) It comprised:

binary weighting scheme:

$$tf(t_i, d_j) = \begin{cases} 1 & \text{if } t_i \in d_j \\ 0 & \text{if } t_i \notin d_j \end{cases}$$

(Absolute) Term-Frequency (TF) weighting scheme:

$$tf(t_i, d_j) = \sum t_i$$

With t_i occurences in d_j.

Document based PCA with binary weight in Figure 9 shows that there are no good clusters in the two first principal components identified.

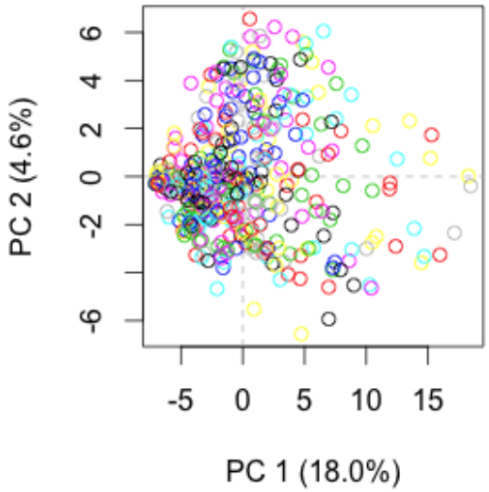


Figure 9: Document based PCA with binary weight

Figure 10 shows the PCA using the terms as outcome variable with binary weight and of a non-sparse matrix. The first component explains approximately a third of the variance compared to the analysis focusing on the documents as it explains only 6.6%. It would require to use around 85 PCAs to explain 90% of the variance, which seems very unsatisfactory given the loss of interpretability.

The score representation of the PCA in Figure 12 shows that the explained variance by the different documents for the component one and two are extremely diverse for the different weights. The (absolute) TF weighting has some clear driving documents, which dominate the whole component. However, this does not help to identify clear clusters or components of documents.

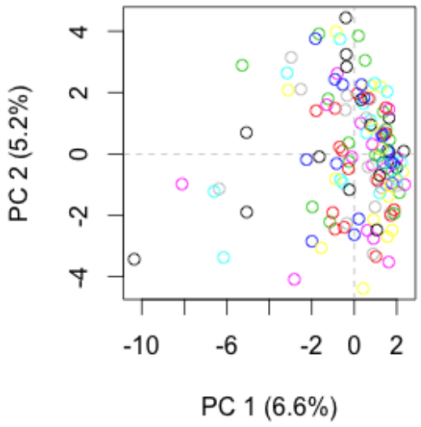


Figure 10: Term based PCA with binary weight

The Figures 11 and 12 show the differences the normalization of TFIDF can have. However, again no clear structures could be identified. Moreover, the loading's of the documents showed some but not necessary helpful dimensionality reductions.

The other model used for the analysis was the Sparse Principal Component Analysis (SPCA). SPCA is closely related to the Principal Component Analysis (PCA), which is a technique for dimensionality reduction. The PCA finds the n-dimensional subspace of maximal variance in the data, which usually only contains non-zero components. The PCA components are a linear combination of the given features. SPCA on the other hand tries to identify a small number of features which still capture a great part of the variance. SPCA enforces sparsity on the components. This leads to a trade-off between sparsity explained variance. However, as in the application of the traditional PCA the SPCA did not produce more useful results. This lead to the decision not to proceed with this approach.

Methods LSA

To improve the findings from Manuscript II, I applied the LSA in the hope to increase the text matching accuracy.

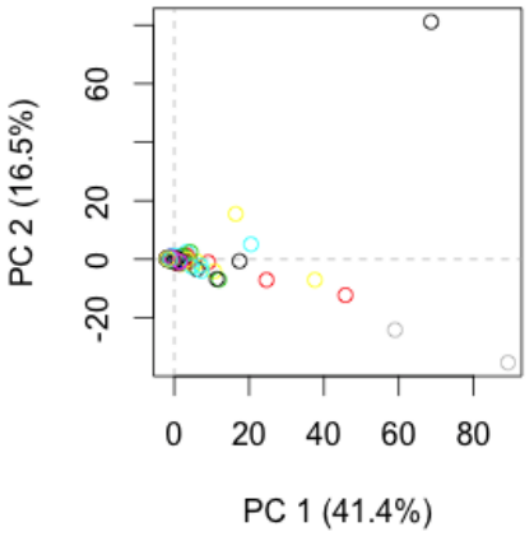


Figure 11: Document based PCA with TF weight

To potentially enhance the findings for hypothesis2 of this study, I used an additional computational method: Latent semantic analysis (LSA). This is another step to improve the detection of text similarities, which is beneficial to ensure an improved matching. LSA is a comparatively old and established method based on single value decomposition (SVD), which provides rank lowering of the DTM (Lee et al., 2010) (see chapter 6). It reduces the highly dimensional the matrix, by creating a new space. LSA decomposes a single matrix into three different spaces: the term representation, document representation and the diagonal matrix.

The DTM of the size is hereby decomposed via a SVD into a matrix displaying term vectors T, a matrix of document vectors D being both orthonormal, and the diagonal matrix S (displaying the singular values). This preserves the ratio among columns. Words or documents are then compared by the cosine of the angle between two vectors. The value 1 represents the most similar even identical words while values close to 0 represent very unrelated words.

$$DTM = TSD^T \tag{2}$$

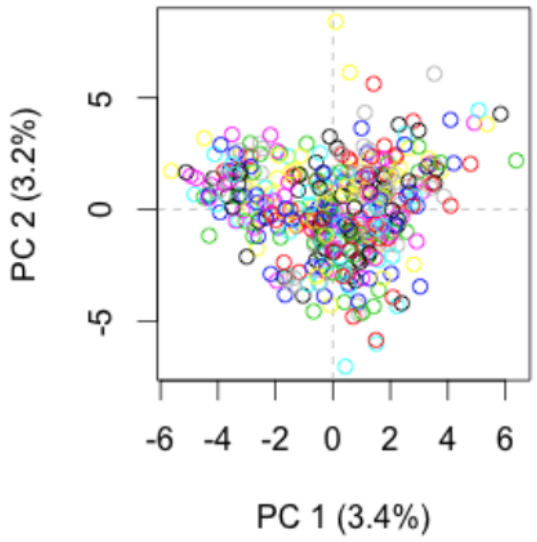


Figure 12: Document based PCA with TFIDF weight

$$LSA = DTM_k \tag{3}$$

$$DTM_k = \sum_{i=1}^{k} T_i \times S_i \times D_i^T \tag{4}$$

K determines the dimensions the LSA space returns. Given the high dimensionality and the sparsity of DTMs the reduction can achieve a much denser representation of the values.

For our purposes, we used an additional *folding in* method, when the common words were $w_n > 100$. Instead of generating a combined LSA space, representing all terms of both documents, we use only words of one matrix (in our case the academic matrices) and fold the other matrix from a website into this pre-existing space. This means, if feasible, only the common term space is taken into consideration, without influencing its factor distribution.

We calculated the cosine between the fold-in matrix space and the original space respectively per column. In the smaller case, for the final computation we generated the document specific space by:

$$LsaDocumentMatrix = S_k \times D_k^T$$

We computed the cosine on this matrix and inspected the top 3 of each corpus comparison. It should be kept in mind that this is based on the LSA matrices, which can lead to any value between −1 and 1, in this case one being the most similar. However, this method is computationally very expensive and is therefore only applied to academic fields with a high probability of knowledge transfer (identified through TFIDF).

Results LSA

Furthermore, we applied the LSA to certain fields for supplementary computational knowledge transfer detection. It was applied it on the most promising academic corpora, such as Photonics. These were identified according to a previous study and on the basis of the TFIDF findings (Woltmann and Alkærsig, 2018).

The cosine similarity varied greatly between the comparisons. Only the cosine similarity measures above 0.8 were stored and prepared for investigation. In some cases, many combinations reached this threshold, so only the highest 3 were taken for later verification to keep the amount of potential matches feasible for human verification. Additionally, only corpora with an overlapping semantic space of a minimum of 100 words (see Section III).

With the application of the LSA, we retrieved very varying potential hits for the first degree partner websites. Some comparisons had extremely high similarity scores, while others were extremely low. However, a manual inspection showed even within the high scoring comparisons no significant content overlap. This suggests that additional measures would be needed to improve the performance of the LSA.

WORD2VEC (W2V)

According to novel and more advanced methods in computational linguistics, I decided to try to train neural network models on the data sets to see whether a better performance for the matching might be possible. The general idea was to extract similar words on the basis of these neural networks and use them for the text matching. W2V are models that are used to represent word embeddings and supposed to identify the linguistic contexts of a word (Mikolov et al., 2013; Rong, 2014). They are two-layer networks that allow to identify similar words for certain contexts[2].

However, trying the application on the data of this thesis did not prove successful. This is mainly due to the limited amount of data. The w2v algorithm needs large amounts of data, such as the entire Google books selection, or other similarly large text material. Hence it was not surprising that the application did not perform well enough in this particular context. However, there might be options to train the models on other data and use them on the thesis data, but this approach could unfortunately not be verified[3].

2 https://skymind.ai/wiki/word2vec

3 More conceptual details and the programming can be obtained up on request-send a message to https://github.com/Salawol

10 DISCUSSION AND CONCLUSION

This chapter summarizes the thesis findings and positions the three attached studies and their findings in relation to each other. The main findings are summarized in the context of the research aim and theoretical, methodological, empirical and policy implications obtained. This chapter discusses the deductive reasoning underlying the implications of my results and highlights some potential practical insights. The chapter is organized in subsections which focus on different aspects of the research implications. It also highlights some technical and theoretical limitations of this research. Finally, it suggests some relevant and technically feasible future research directions.

10.1 THEORETICAL CONTRIBUTION

10.1.1 Conceptual Implications

This section discusses the conclusions derived from this thesis research and the conceptual framework employed. It summarizes some conceptual implications and their relevance for future academic work. I focus mainly on validity and the implications of the research for the current understanding of university-industry Knowledge Transfer, based on an evaluation of the potential of the concepts involved. Not all of the research objectives and studies contribute to our conceptual understanding to the same degree; only the most relevant insights are highlighted here.

The first research objective (RO1 (*"Identify structures and changes to university research knowledge output"*) was addressed by statistical investigation of the different university output dimensions. The university is understood as the producer and active distributor (sender) of knowledge (see chapter 3), and constitutes the basis for our understanding of the proposed input-output model (c.f. chapter 3). The results show that the university output structures constitute a comparatively complex system which undergoes significant changes over time (c.f. Manuscript I). To capture these changes, RO1 investigates research output and its systemic nature (including various types of written output, e.g. publications and patents). I use the output structure and my understanding of it as inputs to the university-industry Knowledge Transfer input-output model. The results from the first study provide three main insights into the input aspect of the model.

First, there are different input dimensions which potentially are relevant to transfer such as accessibility, type (journal publication, etc.) and research field. Second, the input structure changes significantly over time. Third, not all input is comparable in terms of relevance and ratio. These findings have implications for the model input and show that it is crucial to understand the input side of transfer in order to draw conclusions about the process as a whole. Moreover, it can be argued that the composition of the input shapes the outcome of the transfer.

It is evident that the structures in which explicit knowledge is stored and/or disseminated can have a major impact on its final usage. Given these insights into the input dimension, empirical work in this area would be improved by including the input as an explicit aspect. There is clear potential from adding new notions about the relevance of the data composition, and this might increase our understanding of university output in general. Based on the findings in Manuscript I, the integration of additional data such as full text published documents (of whatever type), patents, and technical and consultancy reports would be useful to assess the true amount of research outcomes. This would contribute to a more complete picture of university impact which should be considered in relation to the potential input made by

a contribution. Hence, new considerations regarding existing indicators and data types in terms of their use could be beneficial for university-industry Knowledge Transfer also in larger studies.

In generally, the input-output model proposed in this thesis has proved useful to detect and measure Knowledge Transfer in a novel manner. It could be argued that reducing complex interactions merely to their input and output does not fully reflect the complexity of the transfer mechanisms. However, to measure Knowledge Transfer the proposed model is valid and efficient. In relation to RO2a (*"Develop/adapt computational methods to detect university-industry Knowledge Transfer"*) I was able to demonstrate that observations on both sides of the model are possible and imperative. Identifying Knowledge Transfer can be difficult in an input-output manner where both sides need to be considered and observed. It is comparatively simple to capture research output (model input) but much harder to trace its dissemination on the other side of the model. However, my findings show that it is possible. For example a) the receiving end is observable (although not in its entirety), b) companies can successfully be connected to university research, and c) the outcome might be different from what is predicted in the empirical and other work. I show that new insights can be achieved from a more complete approach which considers both ends of a communication (Shannon, 1948).

By expanding the input-output model with the addition of a dissemination or Knowledge transmission mode, I provide an enhanced model allowing observation of an additional dimension (see Manuscript III) following RO3 (*"Use the methods to investigate potentially relevant dimensions of university-industry Knowledge Transfer"*). In relation to the application of the broadcast and multicast model, the findings in Manuscript III show its potential explanatory power. Here, potential transfer restrictions of a Knowledge Transfer medium (e.g. publications that are Open Access or not) are considered. This approach captures the implications of Open Access. Based on the findings from the third study, it seems that transfer mode (in this case broadcast or multicast) plays a role in effective transfer of knowledge. Although

the interrelation between transfer mode and the transfer outcome is evident, it could be argued that this model does not explain where the differences in specific lie. However, the conceptual approach can be regarded as successful and as offering a new perspective on the potential reach of knowledge transferred from universities to industry. The expanded input-output aligned model based on the adaptation of some simple notions from communication theory (Shannon, 1948) and computer networks (Nevase, 2016) provides information on some additional explanatory aspects which underpin my empirical effort. This basic notion of communication in networks provides adequate and testable explanations which seem to hold for the circumstances tested. Open Access publishing to achieve Knowledge Transfer has clear implications based on the model developed. This thesis provides the foundations for a novel investigation and conceptual framework which is routed in two interrelated scientific disciplines (information systems and communication theory). This interrelation allows further development of explanatory concepts related to Knowledge Transfer which in turn, could provide a common theoretical framework which so far has been lacking (Gherardini and Nucciotti, 2017).

This thesis research supports several well-established understandings, assumptions and notions in the university-industry Knowledge Transfer literature. The findings from the present research show clearly that proximity and/or a relationship such as a formal collaboration between a university and a company lead to a higher probability of Knowledge Transfer (see Manuscript III). This not only supports previous findings from other empirical studies but is in line with the theoretical assumptions about Knowledge Transfer (Drucker and Goldstein, 2007; Siegel et al., 2003). Additionally, it validates two strategic assumptions in this thesis: Since Knowledge Transfer is dependent on formal relationships and proximity in general, it was decided to select formally collaborating and Danish registered firms to obtain a proof of concept.

Moreover, the results of all three studies show that the research field and the age of the research have an impact on how

much the knowledge is disseminated and used by the private sector. Hence, composition and output frequency matter for understanding university research dissemination which is largely in line with previous research (Bekkers and Bodas Freitas, 2008). Overall, I can confirm some current notions about Knowledge Transfer from a different angle and measured in a novel way. I find that in the context of Knowledge Transfer proximity and formal collaboration matter, fields and sectors are important, and types of (publication) output matter. This confirmation of established notions on university-industry Knowledge Transfer can be evaluated to confirm robustness of the method applied. Overall, this thesis improves the theoretical concepts related to bilateral directed exchanges between universities and industry.

10.1.2 Methodological Implications

By following the first research objective (RO1: *"Identify structures and changes to university research knowledge outputs"*) I obtained insights into the dissemination of university research - in particular, the knowledge output system. The findings demonstrate the changes that occur within the publications structures of universities. The methods used suggest that these changes become increasingly relevant in terms of data collection and future empirical work. Also relevant is time; I found that the composition of research output changes over the chosen time frame (2005-2015) along several different dimensions. According to the prior literature, gaps and shortcomings of the traditionally established data sets, changes could be expected, but its extend seems interestingly high.

Although the impact of this finding might not be as dramatic in the case of data sets used in large scale studies, they are relevant whether is concerns a single case study or a large scale international study. In Manuscript I several limitations regarding the different dimensions of university output became evident. The most important ones include the often limited coverage of official publications databases, and the changes in the types of

publications which result in limited data on new and increasing output types (such as e.g. newspaper articles, online blogs (net publications) etc.). The proportions of these changes since the late 2000s show clear trends towards new data structures and shifting relevance of university publishing and output. Therefore methodological advancements such as those aimed for in this thesis, are necessary to prevent the gap in data coverage from becoming wider. Data only on traditional university research outputs might not capture all of a university's scientific research. Better data with wider coverage is needed to identify Knowledge Transfer. In this thesis, I use the university's own database in preference to one of the available official publication databases (c.f. Manuscript II and Manuscript III). This is an important aspect of the second and third studies which investigate research objectives RO2a and RO3.

The attempts made to increase data coverage in Manuscript I proved valuable. Given the large number of bibliometric studies, understanding data shortcomings and potential improvements is useful. Apart from the findings on the coverage and composition of research output, this thesis provides some additional methodological insights in on the basis of the two rather technical research objectives: RO2a: *"Develop/adapt computational methods to detect university-industry Knowledge Transfer"* and RO2b: *"Evaluate the potential of these methods in terms of their potential."* These relate to concrete methods, adaptions and evaluation. RO2a involved tracing university-industry Knowledge Transfer through publications and websites, which was possible but not some challenges. One of the main findings from the proposed text mining methods was that their application to diverse documents types (websites and academic texts) can be problematic in relation to the linguistic differences and formal composition (especially websites). This thesis research highlights the potential and limitations of some tools used in the computational linguistic community.

Chapter 6 investigates these, this community provides many applicable tools using machine learning to investigate text similarities. However, in my case, the complexity of the tasks was too great for their direct application (Rus et al., 2013). Previous

work on established methods and their potential and areas of application (see 6.2.1.) confirm this finding. Study two showed that identifying identical research is difficult but possible. However, the current hype related to machine learning tends to raise expectations to unrealistic levels [1]; Hence, most of these methods fall short of expectations. Simpler algebraic methods were used to compare document types based on creating keyword lists which helped to identify identical content even in the case of text pairs as linguistically different as websites and publications. Extensive use of acronyms and proper nouns in these text types made this task technically possible.

The investigation involved the paradox where the limitations of the documents, short abstracts and short texts on websites which provided too limited data input while others were very long including a few very extensive websites and full-text publications. The long texts of trials and these increased the data load making it time consuming and inefficient to identify matches. Therefore, it can be seen that the application and adaptation of methods is driven by several trade-offs among which the optimum settings and data input (see Manuscript II).

When evaluating the new methodological approach (RO2b), it was clear that technically it is easier to identify common topics and overlapping content than to detect identical research items. This insight is crucial for the academic community since topic modeling is used extensively and often perceived as sufficient for knowledge overlap measure.

Based on the findings from my studies it can be argued that although it might be interesting to discover whether industry and university are working in the same domains or on the same topics, this does not add to the detection of use of common scientific knowledge. Only exact matching provides a measure of Knowledge Transfer and is not replaceable by unsupervised probability prediction (LDA and related topic models (Blei et al., 2003)). Generally the evidence from this research leads to three important

1 https://medium.com/machine-learning-in-practice/business-challenges-with-machine-learning-3d12a32dfd61

insights related to the methods used: First, adaptation of the parameters can be challenging and requires trials, knowledge of the documents and corpora, extensive cleaning (pre-processing), keyword filtering and complex manual verification. These adaptations can improve the results significantly. Keyword detection is difficult in the context of the document types examined in this thesis but the results when certain patterns became evident. Much of the technical pre-processing and threshold setting was done manually but could probably be automated. Second, simple algebraic measures have been shown to provide the best cost-benefit ratios, i.e. least (manual) effort to achieve the best match retrieval. The longstanding and established methods have been shown to outperform advanced (probabilistic) algorithms in this task. This is not surprising due to the high complexity of the real life task which is not easy to decode for the more complex mechanisms. Third, the range of academic fields and websites covered by the data improve the likelihood of identifying matches but makes the task more difficult by requiring additional thresholds. Hence, a more limited (data) approach would have reduced the manual efforts involved, and possibly the errors but might have placed too much restriction on the findings.

These three technical aspects can be considered an evaluation of novel methods, and show that certain aspects need to be considered when applying those methods. In particular, task complexity is especially evident in relation to human verification (see Section III.3.2) and assessment. The findings from the second and third studies suggest that the overall performance of the new methods might not yet be comparable to traditional measures but provides a new perspective on Knowledge Transfer detection. The main methodological implications from this thesis research is that text mining is feasible but needs to be continuously considered along with interpretability of the results.

In evaluating the approach in terms of its potential to provide a novel indicator I can confidently say that this was achieved: The findings can be actually regarded as the creation of an additional indicator for university-industry Knowledge Transfer. This indicator of university-industry Knowledge Transfer is based

on both new and traditional data sources; however, compared to co-patenting and co-publishing (J. Thursby and M. Thursby, 2000), it might not provide the same depth of information. In other words, it does not identify the commercial value of an invention but provides other insights, for instance that a company truly 'knows' about a matter when it publishes the content in the public sphere. Moreover, the data sources used, such as company websites are only proxies for company knowledge but provide insights that other formalized outputs do not capture.

Although the final research objective (RO3: *"Use the methods to investigate potentially relevant dimensions of university-industry Knowledge Transfer"*) is not related directly to methodological aspects it has some relevant implications. It confirms that features of Knowledge Transfer such as Open Access can be examined using the text mining method applied. The findings in Manuscript III demonstrate the potential of this method and provide an understanding of the relevance of certain publications. Overall, despite some technical challenges this first attempt to use computational methods for this type of research has been fruitful for providing additional empirical indicators, measurements and detection of Knowledge Transfer, and demonstrates the potential of computational methods applied to a complex interdisciplinary context.

10.2 POLICY IMPLICATIONS

Investigation of the first research objective (RO1: *"Identify structures and changes to university research knowledge output."*) provides evidence that certain policy changes have clear implications for the composition of explicit knowledge output from universities. These involve not only national policy changes but also university internal changes and notions which have an impact on the composition, ratio and frequency of written output and publication structures.

The two most obvious changes that emerged are the general increase in novel outlets such as internet publications, videos and newspaper articles, and the increase in Open Access publishing. The latter can be attributed to the changing tenor of debates and the political incentives described in chapter 2.1. Therefore, the findings from the first study allow me to conclude that university publication practices adapt to policy decisions and are reflected by the indicator frame targeted.

Investigating the second research objective (RO2a: *"Develop/adapt computational methods to detect university-industry Knowledge Transfer"*; RO2b: *"Evaluate these methods in terms of their potential" "*) generated more insights into policy incentives. The findings show that knowledge is transferred from university to industry, and this drives policy by legitimizing public expenditure (for more detail see chapter 2), and is traceable and measurable to an extent. It allows evaluation of dissemination policies in relation to the effect of the knowledge transferred. This provides insights into the quality/ validity of current traditional proxy-indicators, and could become the basis for an additional indicator to provide a more holistic picture. The insights from this research could be used to make strategic adjustments to policy. For instance, it is important to discover whether formal collaborations that result in commercially relevant outcomes such as co-publication and/or patenting activities indicate also with this novel measure Knowledge Transfer. If so, current strategic goals and indicators are well aligned. If not, then there is a need for other forms of knowledge exchange are as crucial, for instance informal collaborations, and should therefore become more integrated into current policy considerations.

The last research objective (RO3: *"Use the methods to investigate potentially relevant dimensions of university-industry Knowledge Transfer"*) focused on a particular policy incentive: Open Access publishing. Previous studies measure the effectiveness of Open Access in relation to academic contributions, while the importance of Open Access for the industry is often only addressed theoretical (Picarra et al., 2015). This focus and theoretical assumptions have raised expectations about the benefits of Open

Access publishing for society and industry. Although I found clear support for the prediction that easier access enhances Knowledge Transfer, I found that it was less significant and extensive than many policy makers, universities and scholars propose. My results show the need for a reconsideration of these high expectations and assumptions, and investment in more empirical research to validate them. Overall, it seems that Open Access promotes university-industry knowledge dissemination.

It should be noted that the findings related to RO1 show that the time window for this observation and validation might close in near future, and retrospective assessment may become impossible for two reasons. First, the output structure of universities is clearly skewed towards Open Access and policy incentives will ensure that (at least in the Danish case) by 2022 all university publications will be Open Access. This would leave no sample of subscription based publications for comparison with a comparable age. Therefore, as Manuscript III shows, publication age influences the transfer of its content. Hence, to detect Knowledge Transfer differences between the two types requires study of similar time frames. Moreover, online data collection although it could be retroactive, would involve limited data which would hamper the method and not allow direct comparisons. Hence investigation of the potential impact of Open Access publications in terms of knowledge dissemination and transfer must be conducted in the very near future since data sets apart from academic databases will start to become unavailable or unusable for this approach.

Overall, this thesis adds to our understanding of policy effects, implications and new potential assessment strategies. Our findings improve our general understanding of the benefits related to research policy decisions, and raise new questions about current university and public funding incentive structures.

10.3 LIMITATIONS AND FUTURE PERSPECTIVES

This section discusses some limitations of this work and how they could be addressed in future research. It is organized in three subsections according to topics and conceptual aspects. I highlight shortcomings related to the basic underlying assumptions of the thesis; the main aspects of the research strategy used including potential supplementary strategic approaches; and technical limitations related to the statistical methods and tools used.

10.3.1 Conceptual Limitations and Potential Adaptations

Although the conceptual basis of this thesis (i.e. an input-output model based on communications theory notions (see chapter 3)) has proven suitable for the context of this thesis (cf chapter 10.1.1), even though it is a simplified model of reality. This basic simplified assumption has some important limitations and weaknesses. Most notably, such a conceptual foundation includes notions about the sender and receiver units but is not very detailed or multidimensional.

Also, the model is limited by its simplified understanding of the transfer process. It focuses on input and output and treats the intermediate phase as a black box. Accordingly, it does not distinguish between formal and informal Knowledge Transfer (Arundel, 2008) which limits the insights. If the type of channel remains unidentified, understanding the effectiveness and implications of the Knowledge Transfer will be limited. In addition, the interrelations and further exchanges between the two components (university and industry) are not integrated although mutual exchanges of knowledge would seem likely.

The extension to the model which based on the computer networking literature (the broadcast and multicast approach (Nevase, 2016)) is adequate to understand another dimension of Know-

ledge Transfer: transfer channel and relevance of transfer media. However, it provides neither a holistic nor a representative view of reality since it ignores several relevant features such as externalities. These could include government and political frameworks referred to in the triple helix model and their inclusion might be beneficial because they clearly shape the entire exchange by promoting certain transfer channels and modes. Hence, some aspects that have a strong influence on university-industry Knowledge Transfer might have been systematically excluded from the model.

Overall, the model and its extension provide a simplistic basis for an investigation of university-industry Knowledge Transfer. Future research could include the features mentioned above which could be taken from the large literature on channels and externalities. This would help to capture Knowledge Transfer more completely, and add significantly to our understanding of the underlying concepts.

10.3.2 Strategic Choices and Potential Changes

I developed the research strategy based on several considerations and careful evaluation of the possibilities. Since the choice of research strategy involves a trade off, this thesis research has some limitations related to the research design and a) the methodological approach, b) the scope, (including unit(s) of investigation), and c) the indicators and data source(s) considered.

First, I adopted a purely quantitative approach (see chapter 4), which had several advantages and is in line with contemporary empirical studies of Knowledge Transfer detection (see chapter 3). However, there are many qualitative studies of university-industry Knowledge Transfer. Qualitative methods might have shed more light on some aspects and validated some of my assumptions. A more qualitative approach would provide the empirical basis for an examination of certain aspects not investigated empirically in this thesis. For instance, it would be

interesting in future research to investigate the motivations and information seeking strategies of individual industry researchers to understand their use of external knowledge. Breaking the model down into its individual components, and analyzing them qualitatively might add to our understanding of Knowledge Transfer mechanisms and channels.

Second, the strategic decision on scope was important; it limited the scope to two main dimensions in terms of the units of investigation. The focus is on one Scandinavian technical university, and the companies studied had to have at least an indirect relationship to this university. Thus, I test my method on only one case which might mean that the findings are an artifact of this specific case and are not representative of the general reality. A multi-national study would certainly provide more generalizable findings. However, my choice ensured that the scope was manageable for developing and evaluating the novel approach. Future research could use company level data to obtain more information on potential beneficiaries and their features. Furthermore, the defined scope included all the research fields in the university, and all industry sectors. Instead of focusing on a cross section of all research fields and industry sectors, it might have been beneficial to choose specific fields and investigate them in across multiple universities to generate more in-depth insights into single relevant fields or sectors. However, this would have been risky since lack of information on relevant fields might have hampering the success of the research. However, future work could build on this research and study the most relevant fields identified in this thesis.

Third, the choice of data and data sources does not provide a holistic representation of the topic investigated, and hence may influence the final results. The limitations include the representations of knowledge and their coverage in relation to generalizability. Choosing texts to represent knowledge excludes the investigation of the related tacit aspect of knowledge since by definition (), only explicit knowledge can be captured and/or transferred in a written manner. This limits the scope of the present thesis to one aspect of knowledge which does not comprise all the

relevant parameters to assess the entirety of Knowledge Transfer. The tracing and measurement of tacit knowledge remains a gap in the literature.

Moreover, identification of company knowledge through websites is an incomplete and indirect proxy-indicator. Although the reasons for this choice are clear (see chapter 4), it remains a limitation of this thesis. If a technique, software or other research outcome is mentioned on the company's website then it is clear the firm is aware of it but its value to the company is not clear. Hence, it would be crucial to identify the (perceived) value to add the websites to the list of commercially relevant indicators. To understand the actual value of knowledge on websites would be highly beneficial for future research relying on this data type.

Further, a deeper investigation of university-industry knowledge would require knowledge related data provided by one or two large companies; this could provide insights not captured by observing public websites. This would shed additional light on university-industry Knowledge Transfer and its potential. Overall, the possibilities identified for future research would provide relevant information and clearly seem feasible.

10.3.3 Method Limitations and Technical Insights

The choice of text mining methods and data selection and collection limit the choice of potential statistical tools and algorithms applied. The general limitations of text mining as the chosen method include: a) the understanding of text content and synonymy of terms is still limited; b) corpus specific cleaning and pre-processing is difficult but decisive for outcomes; c) linguistic composition matters and cannot be influenced; d) language is important since cross lingual approaches are neither common nor (yet) successful. This research has some limitations related to the choice of machine learning algorithms (e.g. topic modeling) in particular with the choice of unsupervised approaches. Although an unsupervised approach was needed due to the infeasibility of

labeling possibilities, the results are sparse and could certainly be improved by application of semi-supervised or supervised approaches. This limitation applies to many unsupervised machine learning algorithms which makes them less valuable for certain classifications or similarity measures.

In the case of the present research, data availability was a constraint in the context of these methods since the academic abstracts were often too short for advanced assessment. Also, deep learning or the application of neural networks was not feasible for the "small" data set. At the same time, some large websites involve a lot of noise which additionally limits performance. It is doubtful that if I would have used all the website data available I could have improved the outcomes. The small number of verified results for companies with no formal relationships would not have added any value (see Manuscript III). This could hold for several studies based on machine learning where the amount of data seems more valuable than the overview about the actual turn outs. Given this study's empirical scope it seems that the chosen setting was useful for a first proof of concept for the following reasons:

1. Data amounts were manageable and the necessary manual examinations were possible;

2. English being the publication and (often) corporate language in Denmark was a technical advantage;

3. The focus on applied sciences in the technical university increased the likeliness of verified matches;

4. The extensive collaboration of this university with industry provided a large list of potential knowledge receivers also increased occurrence likelihood.

Taken together, these aspects provided a good basis for investigating the research question and the research objectives in a strategic and meaningful manner while also highlighting some limitations and potential enhancements in future research. Many

of the limitations identified can be addressed by the computational linguistics and machine learning community in the next years.

Hence, I expect that some of the problems will be reduced by use of future developments in the machine learning community such as "transfer learning". Transfer learning generates and trains machine learning algorithms on data sets and transfers them to other data with the need for only minor changes. It is proving successful in several domains including image analysis. The algebraic methods I used have some shortcomings including that the indexing of important words is very localized and may miss important global words. This is problematic in the case of academic abstracts in particular. For instance TFIDF indexing removes the word "wind" from wind energy topics which obviously eliminates a content relevant term. This could for instance be resolved by applying an additional measure at corpus level and combine them with the document specific keyword lists.

Manual verification is not straightforward, suggesting that the chosen data which were necessary for the thesis scope were technically very ambitious and advanced. Simpler texts would be easier to classify and identify similarities but do not provide insights into university-industry Knowledge Transfer. Nevertheless, the performance of text mining tools depends on task complexity. Further work in this direction will require additional computational validation possibilities. There are some statistical problems and implications related to the application of different data types and data sets. The composition of the various data sets raises issues ranging from data extraction and data coverage to identification of the relevant variables etc. Not all of these processes were initially successful; some significant adaptations were needed.

10.4 EVALUATION OF RESEARCH AIM

I investigated whether current understanding of university-industry Knowledge Transfer could be enhanced by the application of novel computational methods. I can generally conclude that the application of novel methods offers novel possibilities which was the aim of this thesis research. From a methodological point of view the thesis objectives have been achieved although there are some unresolved issues related in particular to feasibility and performance of the methods.

However, developments in machine learning and similar computational fields should be able to solve these issues in the short term. For stakeholders interested in university-industry Knowledge Transfer, i.e. politicians, academic researchers and society, I identified new possibilities to trace and validate Knowledge Transfer and, accordingly, validate and improve current transfer strategies and activities. This can help legitimizing the public expenditure and potentially even increase the impact of university research. It could shape and improve transfer activities through adaptations to outcome directed policies and evidence-based incentive structures to ensure reliable results. This could foster innovation and development in national settings.

To conclude, increasing understanding and improving measurements are not easy. However, I consider that this thesis is a first step in this direction and opens up new research opportunities that could guide future notions about university impact.

BIBLIOGRAPHY

Agrawal, Ajay (2001). 'University-to-industry knowledge transfer: Literature review and unanswered questions'. In: *International Journal of management reviews* 3.4, pp. 285–302.

Agrawal, Ajay and Rebecca Henderson (2002). 'Putting Patents in Context: Exploring Knowledge Transfer from MIT'. In: *Mgmt. Sci.* 48.1, pp. 44–60. arXiv: Signatur:N.axy.

Ankrah, Samuel N, Thomas F Burgess, Paul Grimshaw and Nicky E Shaw (2013). 'Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit'. In: *Technovation* 33.2-3, pp. 50–65.

Antelman, Kristin (2004). 'Do open-access articles have a greater research impact?' In: *College & research libraries* 65.5, pp. 372–382.

Argote, Linda and Paul Ingram (2000). 'Knowledge Transfer : A Basis for Competitive Advantage in Firms'. In: *Organizational Behavior and Human Decision Processes* 82.1, pp. 150–169.

Armstrong, Mark (2015). 'Opening access to research'. In: *The Economic Journal* 125.586, F1–F30.

Arundel Anthony Bordoy, Catalina (2008). 'Developing internationally comparable indicators for the commercialization of publicly-funded research'. In: *Maastricht: UNU-MERIT* 31, pp. 1–32.

Auranen, Otto and Mika Nieminen (2010). 'University research funding and publication performance—An international comparison'. In: *Research Policy* 39.6, pp. 822–834.

Balconi, Margherita and Andrea Laboranti (2006). 'University–industry interactions in applied research: The case of microelectronics'. In: *Research Policy* 35.10, pp. 1616–1630.

Banjade, Rajendra, Nabin Maharjan, Nobal B Niraula, Vasile Rus and Dipesh Gautam (2015). 'Lemon and tea are not similar: Measuring word-to-word similarity by combining different

methods'. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 335–346.

Banjade, Rajendra, Nabin Maharjan, Nobal Bikram Niraula and Vasile Rus (2016). 'DTSim at SemEval-2016 Task 2: Interpreting Similarity of Texts Based on Automated Chunking, Chunk Alignment and Semantic Relation Prediction'. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 809–813.

Bekkers, Rudi and Isabel Maria Bodas Freitas (2008). 'Analysing knowledge transfer channels between universities and industry: To what degree do sectors also matter?' In: *Research Policy* 37.10, pp. 1837–1853.

Bentley, Peter James, Magnus Gulbrandsen and Svein Kyvik (2015). 'The relationship between basic and applied research in universities'. In: *Higher Education* 70.4, pp. 689–709.

Bercovitz, Janet and Maryann Feldman (2006). 'Entpreprenerial universities and technology transfer: A conceptual framework for understanding knowledge-based economic development'. In: *The Journal of Technology Transfer* 31.1, pp. 175–188.

Berry, Michael W and Malu Castellanos (2004). 'Survey of text mining'. In: *Computing Reviews* 45.9, p. 548.

Berry, Michael W and Malu Castellanos (2007). 'Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition'. In: p. 241.

Bird, Steven, Ewan Klein and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Bishop, Kate, Pablo D'Este and Andy Neely (2011). 'Gaining from interactions with universities: Multiple methods for nurturing absorptive capacity'. In: *Research Policy* 40.1, pp. 30–40.

Blei, David M, Andrew Y Ng and Michael I Jordan (2003). 'Latent dirichlet allocation'. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bornmann, Lutz (2013). 'What is societal impact of research and how can it be assessed? A literature survey'. In: *Journal of the American Society for Information Science and Technology* 64.2, pp. 217–233.

Bozeman, Barry (2000). 'Technology transfer and public policy: a review of research and theory'. In: *Research policy* 29.4-5, pp. 627–655.

Brennenraedts, Reginald, Rudi Bekkers, Bart Verspagen et al. (2006). 'The different channels of university-industry knowledge transfer: Empirical evidence from Biomedical Engineering'. In: *Eindhoven, Eindhoven Centre for Innovation Studies, The Netherlands*.

Brostroem, Anders (2012). 'Firms' rationales for interaction with research universities and the principles for public co-funding'. In: *Journal of Technology Transfer* 37.3, pp. 313–329.

Bruneel, Johan, Pablo D'Este and Ammon Salter (2010). 'Investigating the factors that diminish the barriers to university–industry collaboration'. In: *Research policy* 39.7, pp. 858–868.

Bryman, Alan (2012). *Social Research Methods*. Oxford University Press.

Calvert, Jane and Pari Patel (2003). 'University-industry research collaborations in the UK: bibliometric trends'. In: *Science and public policy* 30.2, pp. 85–96.

Cattaneo, Mattia, Michele Meoli and Andrea Signori (2016). 'Performance-based funding and university research productivity: The moderating effect of university legitimacy'. In: *The Journal of Technology Transfer* 41.1, pp. 85–104.

Cheah, Sarah (2016). 'Framework for measuring research and innovation impact'. In: *Innovation* 18.2, pp. 212–232.

Cohen, Wesley and Daniel A. Levinthal (1990). 'Absorptive Capacity: A New Perspective on Learning and Innovation'. In: *Administrative Science Quarterly* 35.1, pp. 128–152.

Cohen, Wesley, Richard R Nelson and John P Walsh (2002). 'Links and impacts: the influence of public research on industrial R&D'. In: *Management science* 48.1, pp. 1–23.

Crespi, Gustavo, Pablo D'Este, Roberto Fontana and Aldo Geuna (2011). 'The impact of academic patenting on university research and its transfer'. In: *Research policy* 40.1, pp. 55–68.

Creswell, John W (1994). 'Research design: Qualitative, quantitative, and mixed methods approaches'. In:

Danish Council for Independent Research, the Danish National Research Foundation, Danish Council for Strategic Research and Danish National Advanced Technology Foundation and

the Danish Council (2012). 'Open Access policy for public research councils and foundations'. In:

De Long, David W and Liam Fahey (2000). 'Diagnosing cultural barriers to knowledge management'. In: *Academy of Management Perspectives* 14.4, pp. 113–127.

D'Este, Pablo and Parimal Patel (2007). 'University–industry linkages in the UK: What are the factors underlying the variety of interactions with industry?' In: *Research policy* 36.9, pp. 1295–1313.

D'Este, Pablo and Markus Perkmann (2011). 'Why do academics engage with industry? The entrepreneurial university and individual motivations'. In: *The Journal of Technology Transfer* 36.3, pp. 316–339.

D'Este, Pablo, Irene Ramos-Vielba, Richard Woolley and Nabil Amara (2018). 'How do researchers generate scientific and societal impacts? Toward an analytical and operational framework'. In: *Science and Public Policy*.

Dosi, Giovanni and Richard R Nelson (1994). 'An introduction to evolutionary theories in economics'. In: *Journal of evolutionary economics* 4.3, pp. 153–172.

Drucker, Joshua and Harvey Goldstein (2007). 'Assessing the regional economic development impacts of universities: A review of current approaches'. In: *International regional science review* 30.1, pp. 20–46.

Etzkowitz, Henry (2008). *The triple helix: university-industry-government innovation in action*. Routledge.

Etzkowitz, Henry and Loet Leydesdorff (1995). 'The Triple Helix–University-industry-government relations: A laboratory for knowledge based economic development'. In:

Etzkowitz, Henry, Andrew Webster, Christiane Gebhardt and Branca Regina Cantisano Terra (2000). 'The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm'. In: *Research policy* 29.2, pp. 313–330.

Feldman, Ronen, James Sanger et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Feller, I (1990). 'Universities as engines of R& D-based economic growth: They think they can'. In: *Research Policy* 19.4, pp. 335–348.

Franco, Mário and Heiko Haase (2015). 'University–industry cooperation: Researchers' motivations and interaction channels'. In: *Journal of Engineering and Technology Management* 36, pp. 41–51.

Fung, Hon-Ngen and Chan-Yuan Wong (2017). 'Scientific collaboration in indigenous knowledge in context: Insights from publication and co-publication network analysis'. In: *Technological Forecasting and Social Change* 117, pp. 57–69.

Gabrilovich, Evgeniy and Shaul Markovitch (2007). 'Computing semantic relatedness using wikipedia-based explicit semantic analysis.' In: *IJcAI*. Vol. 7, pp. 1606–1611.

Gettier, Edmund L (1963). 'Is justified true belief knowledge?' In: *analysis* 23.6, pp. 121–123.

Geuna, Aldo (2001). 'The changing rationale for European university research funding: are there negative unintended consequences?' In: *Journal of economic issues* 35.3, pp. 607–632.

Geuna, Aldo and Ben R Martin (2003). 'University research evaluation and funding: An international comparison'. In: *Minerva* 41.4, pp. 277–304.

Gherardini, Alberto and Alberto Nucciotti (2017). 'Yesterday's giants and invisible colleges of today. A study on the 'knowledge transfer' scientific domain'. In: *Scientometrics* 112.1, pp. 255–271.

Gomaa, Wael H and Aly A Fahmy (2013). 'A survey of text similarity approaches'. In: *International Journal of Computer Applications* 68.13, pp. 13–18.

Grimpe, C and K Hussinger (2013). 'Formal and Informal Knowledge and Technology Transfer from Academia to Industry: Complementarity Effects and Innovation Performance'. In: *Industry and Innovation* 20.January 2016, pp. 683–700.

Grün, Bettina and Kurt Hornik (2011). 'topicmodels : An R Package for Fitting Topic Models'. In: *Journal of Statistical Software* 40.13, pp. 1–30. URL: http://kortan.sote.hu/ftp/mirrors/CRAN/web/packages/topicmodels/vignettes/topicmodels.pdf.

Gulbrandsen, Magnus and Stig Slipersaeter (2007). 'The third mission and the entrepreneurial university model'. In: *Universities and Strategic Knowledge Creation: Specialization and Performance in Europe*. Chap. 4, pp. 112–143. URL: https://books.

google.dk/books?hl=da%7B%5C&%7Dlr=%7B%5C&%7Did=8wlPwc5wMzoC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PA112%7B%5C&%7Ddq=third+mission+and+the+entrepeneural+model%7B%5C&%7Dots=ONqLIqh4o4%7B%5C&%7Dsig=PRqAJCZiFSyZIxhtTE%7B%5C_%7Dh2fsWh5M%7B%5C&%7Dredir%7B%5C_%7Desc=y%7B%5C#%7Dv=onepage%7B%5C&%7Dq=third%20mission%20and%20the%20entrepeneural%20model%7B%5C&%7Df=false.

Henderson, Rebecca, Adam B Jaffe and Manuel Trajtenberg (1998). 'Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988'. In: *Review of Economics and statistics* 80.1, pp. 119–127.

Hicks, Diana (2012). 'Performance-based university research funding systems'. In: *Research Policy* 41.2, pp. 251–261. URL: http://www.sciencedirect.com/science/article/pii/S0048733311001752.

Higher Education Funding Council for England (2016). *Policy for open access in Research Excellence Framework 2021*.

Howells, Jeremy (2002). 'Tacit Knowledge, Innovation and Economic Geography'. In: *Urban Studies* 39.5-6, pp. 871–884.

Howells, Jeremy, Ronnie Ramlogan and Shu-Li Cheng (2012). 'Innovation and university collaboration: paradox and complexity within the knowledge economy'. In: *Cambridge Journal of Economics* 36.3, pp. 703–721.

Hübner, David, Thibault Verhoeven, Konstantin Schmid, Klaus-Robert Müller, Michael Tangermann and Pieter-Jan Kindermans (2017). 'Learning from label proportions in brain-computer interfaces: Online unsupervised learning with guarantees'. In: *PloS one* 12.4, e0175856.

Huggins, Robert and Andrew Johnston (2009). 'The economic and innovation contribution of universities: A regional perspective'. In: *Environment and Planning C: Government and Policy* 27.6, pp. 1088–1106.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jokić, Maja, Andrea Mervar and Stjepan Mateljan (2018). 'Scientific potential of European fully open access journals'. In: *Scientometrics* 114.3, pp. 1373–1394.

Jongbloed, Ben, Jürgen Enders and Carlo Salerno (2008). 'Higher education and its communities: Interconnections, interdependencies and a research agenda'. In: *Higher Education* 56.3, pp. 303–324.

Kayser, Victoria and Knut Blind (2017). 'Extending the knowledge base of foresight: The contribution of text mining'. In: *Technological Forecasting and Social Change* 116, pp. 208–215.

Kenney, Martin and Donald Patton (2009). 'Reconsidering the Bayh-Dole Act and the current university invention ownership model'. In: *Research Policy* 38.9, pp. 1407–1422.

Kinne, Jan and Janna Axenbeck (2018). *Web mining of firm websites: A framework for web scraping and a pilot study for Germany*. Tech. rep. ZEW Discussion Papers.

Kuhlthau, Carol C (1991). 'Inside the search process: Information seeking from the user's perspective'. In: *Journal of the American society for information science* 42.5, pp. 361–371.

Kwon, Ki-Seok, Han Woo Park, Minho So and Loet Leydesdorff (2012). 'Has globalization strengthened South Koreaś national research system? National and international dynamics of the Triple Helix of scientific co-authorship relationships in South Korea'. In: *Scientometrics* 90.1, pp. 163–176.

Laredo, Philippe (2007). 'Revisiting the third mission of universities: Toward a renewed categorization of university activities?' In: *Higher education policy* 20.4, pp. 441–456.

Laursen, Keld and Ammon Salter (2004). 'Searching high and low: What types of firms use universities as a source of innovation?' In: *Research Policy* 33.8, pp. 1201–1215.

Laursen, Keld and Ammon Salter (2014). 'The paradox of openness: Appropriability, external search and collaboration'. In: *Research Policy* 43.5, pp. 867–878.

Lee, Sangno, Jaeki Song and Yongjin Kim (2010). 'An Empirical Comparison of Four Text Mining Methods'. In: *Journal of Computer Information System*, pp. 1–10.

Lengyel, Balázs and Loet Leydesdorff (2011). 'Regional innovation systems in Hungary: The failing synergy at the national level'. In: *Regional Studies* 45.5, pp. 677–693.

Leydesdorff, Loet (2012). 'The Triple Helix of University-Industry-Government Relations (February 2012)'. In:

Leydesdorff, Loet and Henry Etzkowitz (1998). 'The triple helix as a model for innovation studies'. In: *Science and public policy* 25.3, pp. 195–203.

Link, Albert, Donald Siegel and Barry Bozeman (2007). 'An empirical analysis of the propensity of academics to engage in informal university technology transfer'. In: *Industrial and Corporate Change* 16.4, pp. 641–655.

Lissoni, Francesco, Patrick Llerena, Maureen McKelvey and Bulat Sanditov (2008). 'Academic patenting in Europe: new evidence from the KEINS database'. In: *Research Evaluation* 17.2, pp. 87–102.

Lissoni, Francesco, Peter Lotz, Jens Schovsbo and Adele Treccani (2009). 'Academic patenting and the professor's privilege: evidence on Denmark from the KEINS database'. In: *Science and Public Policy* 36.8, pp. 595–607. URL: http://spp.oxfordjournals.org/content/36/8/595.abstract.

Liyanage, C., T. Ballal, T. Elhag and Q. Li (2009). 'Knowledge communication and translation - a knowledge transfer model'. In: *Journal of Knowledge Management* 13.3, pp. 118–131. URL: http://www.emeraldinsight.com/doi/abs/10.1108/13673270910962914.

Martin, Ben R (2011). 'The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?' In: *Research evaluation* 20.3, pp. 247–254.

Meyer, Martin, Kevin Grant, Piera Morlacchi and Dagmara Weckowska (2014). 'Triple Helix indicators as an emergent area of enquiry: a bibliometric perspective'. In: *Scientometrics* 99.1, pp. 151–174.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean (2013). 'Distributed representations of words and phrases and their compositionality'. In: *Advances in neural information processing systems*, pp. 3111–3119.

Miner, Gary, John Elder IV and Thomas Hill (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Ministry of Higher Education (2018). *Danmarks Nationale Strategi for Open Access*.

Moed, Henk F, WJM Burger, JG Frankfort and Anthony FJ Van Raan (1985). 'The use of bibliometric data for the measurement

of university research performance'. In: *Research policy* 14.3, pp. 131–149.

Mowery, David C and Bhaven N Sampat (2005). 'Universities in National Innovation Systems'. In: *The Oxford Handbook of Innovation*.

Mowery, David C, Bhaven N Sampat and Arvids A Ziedonis (2002). 'Learning to patent: Institutional experience, learning, and the characteristics of US university patents after the Bayh-Dole Act, 1981-1992'. In: *Management Science* 48.1, pp. 73–89.

Munari, Federico, Einar Rasmussen, Laura Toschi and Elisa Villani (2016). 'Determinants of the university technology transfer policy-mix: A cross-national analysis of gap-funding instruments'. In: *The Journal of Technology Transfer* 41.6, pp. 1377–1405.

Nelson, Richard R (1992). 'National innovation systems: a retrospective on a study'. In: *Industrial and corporate change* 1.2, pp. 347–374.

Nevase, Viraj (2016). 'A Practical Guide to Differentiate Unicast, Broadcast & Multicast'. In: URL: https://www.esds.co.in/blog/difference-between-unicast-broadcast-and-multicast/.

Nogueira, Bruno M, Maria F Moura, M da S CONRADO, Rafael G Rossi, Ricardo M Marcacini and Solange O Rezende (2008). 'Winning some of the document preprocessing challenges in a text mining process.' In: *Embrapa Informática Agropecuária-Artigo em anais de congresso (ALICE)*. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 23.; SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE, 22.; WORKSHOP EM ALGORITMOS E APLICAÇÕES DE MINERAÇÃO DE DADOS, 4., 2008, Campinas. Anais... Campinas: UNICAMP, Instituto de Computação, 2008.

Olson, David R (1996). *The world on paper: The conceptual and cognitive implications of writing and reading*. Cambridge University Press.

O'Shea, Rory P., Harveen Chugh and Thomas J. Allen (2008). 'Determinants and consequences of university spinoff activity: A conceptual framework'. In: *Journal of Technology Transfer* 33.6, pp. 653–666.

Park, Albert, Mike Conway and Annie T Chen (2018). 'Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach'. In: *Computers in human behavior* 78, pp. 98–112.

Park, Han Woo, Heung Deug Hong and Loet Leydesdorff (2005). 'A comparison of the knowledge-based innovation systems in the economies of South Korea and the Netherlands using Triple Helix indicators'. In: *Scientometrics* 65.1, pp. 3–27.

Partha, Dasgupta and Paul A David (1994). 'Toward a new economics of science'. In: *Research policy* 23.5, pp. 487–521.

Patra, Swapan Kumar, Partha Bhattacharya and Neera Verma (2003). 'Bibliometric study of literature on bibliometrics'. In: *DESIDOC Journal of Library & Information Technology* 26.1.

Paukkeri, MS and Timo Honkela (2010). 'Likey: Unsupervised language-independent keyphrase extraction'. In: *Proceedings of the 5th International Workshop ...* July, pp. 162–165. URL: http://dl.acm.org/citation.cfm?id=1859698.

Pavitt, Keith (1991). 'What makes basic research economically useful?' In: *Research policy* 20.2, pp. 109–119.

Perkmann, Markus, Andy Neely and Kathryn Walsh (2011). 'How should firms evaluate success in university–industry alliances? A performance measurement system'. In: *R&D Management* 41.2, pp. 202–216.

Perkmann, Markus and Kathryn Walsh (2009). 'The two faces of collaboration: Impacts of university-industry relations on public research'. In: *Industrial and Corporate Change* 18.6, pp. 1033–1065.

Perkmann, Markus, Valentina Tartari, Maureen McKelvey, Erkko Autio, Anders Broström, Pablo D'Este, Riccardo Fini, Aldo Geuna, Rosa Grimaldi, Alan Hughes et al. (2013). 'Academic engagement and commercialisation: A review of the literature on university–industry relations'. In: *Research policy* 42.2, pp. 423–442.

Perry, Chad and Oystein Jensen (2001). 'Approaches to combining induction and deduction in one research study'. In: *Conference of the Australian and New Zealand Marketing Academy, Auckland, New Zealand*.

Pesole, Annarosa, Daniel Nepelski et al. (2016). 'Universities and collaborative innovation in EC-funded research projects: An analysis based on Innovation Radar data'. In: *JRC Scientific and Policy Reports–EUR* 28355.

Phillips, Derek L (1974). 'Epistemology and the sociology of knowledge: The contributions of Mannheim, Mills, and Merton'. In: *Theory and Society* 1.1, pp. 59–88.

Picarra, Mafalda, EKT Victoria Tsoukala and Alma Swan (2015). 'Open Access to scientific information: facilitating knowledge transfer and technological innovation from the academic to the private sector'. In: *PASTEUR4OA Briefing Paper*.

Polanyi, Michael (1962). 'Tacit knowing: Its bearing on some problems of philosophy'. In: *Reviews of modern physics* 34.4, p. 601.

Ponweiser, Martin (2012). 'Latent Dirichlet allocation in R'. In:

Ramos-Vielba, Irene, Manuel Fernández-Esquinas and Elena Espinosa-de-los-Monteros (2009). 'Measuring university–industry collaboration in a regional innovation system'. In: *Scientometrics* 84.3, pp. 649–667.

Rasmussen, Einar, Øystein Moen and Magnus Gulbrandsen (2006). 'Initiatives to promote commercialization of university knowledge'. In: *Technovation* 26.4, pp. 518–533.

Richardson, G. Manning, Janet Bowers, a. John Woodill, Joseph R. Barr, Jean Mark Gawron and Richard a. Levine (2014). 'Topic Models: A Tutorial with R'. In: *International Journal of Semantic Computing* 08.01, pp. 85–98. URL: http://www.worldscientific.com/doi/abs/10.1142/S1793351X14500044.

Robson, Colin and Kieran McCartan (2016). *Real world research*. John Wiley & Sons.

Rong, Xin (2014). 'word2vec parameter learning explained'. In: *arXiv preprint arXiv:1411.2738*.

Roper, Stephen, James H Love and Karen Bonner (2017). 'Firms' knowledge search and local knowledge externalities in innovation performance'. In: *Research Policy* 46.1, pp. 43–56.

Rossi, Federica and Ainurul Rosli (2015). 'Indicators of university–industry knowledge transfer performance and their implications for universities: evidence from the United Kingdom'. In: *Studies in Higher Education* 40.10, pp. 1970–1991.

Rothaermel, Frank T., Shanti D. Agung and Lin Jiang (2007). 'University entrepreneurship: A taxonomy of the literature'. In: *Industrial and Corporate Change* 16.4, pp. 691–791.

Rus, Vasile (2014). 'Opportunities and Challenges in Semantic Similarity.' In: *FLAIRS Conference*.

Rus, Vasile, Nobal Niraula and Rajendra Banjade (2013). 'Similarity measures based on latent dirichlet allocation'. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 459–470.

Salter, Ammon and Ben Martin (2001). 'The economic benefits of publicly funded basic research: a critical review'. In: *Res. Policy* 30.3, pp. 509–532. URL: http://www.sciencedirect.com/science/article/pii/S0048733300000913%7B%5C%%7D5CnAll%20Papers/S/Salter%20and%20Martin%202001%20-%20The%20economic%20benefits%20of%20publicly%20funded%20basic%20research%20-%20a%20critical%20review.pdf.

Santoro, Michael D and Paul E Bierly (2006). 'Facilitators of knowledge transfer in university-industry collaborations: A knowledge-based perspective'. In: *IEEE Transactions on engineering management* 53.4, pp. 495–507.

School of Electronics and Computer Science at the University of Southampton England (2018). *Registry of Open Access Repository Mandates and Policies*. URL: https://roarmap.eprints.org/.

Schwarz, A Winkel, Stephan Schwarz and Robert Tijssen (1998). 'Research and research impact of a technical university—A bibliometric study'. In: *Scientometrics* 41.3, pp. 371–388.

Scott, John C (2006). 'The mission of the university: Medieval to postmodern transformations'. In: *The journal of higher education* 77.1, pp. 1–39.

Shannon, Claude Elwood (1948). 'A mathematical theory of communication'. In: *Bell system technical journal* 27.3, pp. 379–423.

Siegel, Donald, Reinhilde Veugelers and Mike Wright (2007). 'Technology transfer offices and commercialization of university intellectual property: performance and policy implications'. In: *Oxford review of economic policy* 23.4, pp. 640–660.

Siegel, Donald, David A. Waldman, Leanne E. Atwater and Albert Link (2004). 'Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: Qualitative evidence from the commercialization of university tech-

nologies'. In: *Journal of Engineering and Technology Management - JET-M* 21.1-2, pp. 115–142. arXiv: arXiv:1011.1669v3.

Siegel, Donald, David Waldman, Leanne E. Atwater and Albert Link (2003). 'Commercial knowledge transfers from universities to firms: Improving the effectiveness of university-industry collaboration'. In: *Journal of High Technology Management Research* 14.1, pp. 111–133.

Sukamolson, Suphat (2007). 'Fundamentals of quantitative research'. In: *Language Institute Chulalongkorn University*, pp. 1–20.

Tennant, Jonathan P, François Waldner, Damien C Jacques, Paola Masuzzo, Lauren B Collister and Chris HJ Hartgerink (2016). 'The academic, economic and societal impacts of Open Access: an evidence-based review'. In: *F1000Research* 5.

Thursby, Jerry, Richard Jensen and Marie Thursby (2001). 'Objectives, characteristics and outcomes of university licensing: A survey of major US universities'. In: *The Journal of Technology Transfer* 26.1, pp. 59–72.

Thursby, Jerry and Marie Thursby (2000). 'The Bayh-Dole Act'. In:

Thursby, Jerry and Marie Thursby (2003). *University licensing and the Bayh-Dole act*.

Wilson, Thomas D (2000). 'Human information behavior'. In: *Informing science* 3.2, pp. 49–56.

Woltmann, Sabrina and Lars Alkærsig (2017). 'Search for Knowledge: Text Mining for Examination of University-Industry Knowledge Transfer'. In: *DRUID17*.

Woltmann, Sabrina and Lars Alkærsig (2018). 'Tracing university–industry knowledge transfer through a text mining approach'. In: *Scientometrics*. URL: https://doi.org/10.1007/s11192-018-2849-9.

Woltmann, Sabrina, Sebastiano Piccolo and Melanie Kreye (2018). 'Understanding The Diversity Of University Research Knowledge Structures And Their Development Over Time'. In: *23rd International Conference on Science and Technology Indicators (STI 2018)*.

Xia, Tian and Yanmei Chai (2011). 'An improvement to TF-IDF: Term distribution based term weight algorithm'. In: *Journal of Software* 6.3, pp. 413–420.

Xu, Guandong, Yanchun Zhang and Lin Li (2010). *Web mining and social networking: techniques and applications*. Vol. 6. Springer Science & Business Media.

Young, Tom, Devamanyu Hazarika, Soujanya Poria and Erik Cambria (2017). 'Recent trends in deep learning based natural language processing'. In: *arXiv preprint arXiv:1708.02709*.

Zagzebski, Linda (2017). 'What is knowledge?' In: *The Blackwell guide to epistemology*, pp. 92–116.

APPENDICES

A CONFERENCE: STI VALENCIA 2016

Title: *From university research to innovation Detecting knowledge transfer via text mining*
Author: Woltmann, Sabrina and Clemmensen, Line H and Alkærsig, Lars
Conference: 21st international conference on Science and Technology Indicators (STI 2016) Science and Technology Indicators Conference
Year: *2016*

STI Conference 2016 · València Peripheries, frontiers and beyond
València, 14 · 16 September 2016

Proceedings of the 21st International Conference on Science and Technology Indicators | València (Spain) | September 14-16, 2016
DOI: http://dx.doi.org/10.4995/STI2016.2016.xxxx

From university research to innovation
Detecting knowledge transfer via text mining

Sabrina Woltmann* Line H. Clemmensen** and Lars Alkærsig*

*swol@dtu.dk; lalk@dtu.dk
Management Engineering, Technical University of Denmark, Centrifugevej 372, 2800 Kgs. Lyngby, (Denmark)

** lkhc@dtu.dk
Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, (Denmark)

INTRODUCTION

Universities are facing increasing demands for active dissemination of their research results and their contribution to knowledge development in their socioeconomic environment (Jongbloed, Enders & Salerno 2008); commonly referred to as the third mission. Since knowledge is a crucial aspect for innovation processes, its transfer has become a new policy priority and is often directly targeted by public funding (Ramos-Vielba, Fernández-Esquinas & Espinosa-de-los-Monteros 2009, Huggins & Johnston 2009). This study covers the extent of knowledge transfer of university research within technical sciences, as these are key drivers for innovation.

Current empirical research focuses primarily on the analysis of formal interactions between universities and their company partners (Broström 2010), relying on indicators such as patents, collaborative publications, contracts and license agreements (Drucker & Goldstein 2007). These well-developed empirical approaches somewhat capture the success of knowledge dissemination and commercialization of university driven innovations. However, these studies bear some deficiencies, as they often fail to include indirect impacts by focusing on formal cooperation and knowledge exchange. Additionally, most empirical studies frequently require complex adjustments for each unique case. Moreover, their key indicators often depend on formal databases with varying quality and accessibility and they require long-term assessments, which delays the outcomes and limit comparability (Vincett 2010).

In this study, we use modern computational methods to expand the empirical framework by introducing specific data mining approaches and testing these on the Technical University of Denmark (DTU). To complement the current scope, we focus in particular on the application of text mining and pattern recognition tools. These tools capture occurrences where knowledge is used without a statement about its origin. Our data sources include the online presence of companies in regional proximity to the university, including social media sites, company websites and annual reports.

The study's intent is to counteract certain empirical challenges, by detecting knowledge transfer without focusing on formal cooperation channels and develop additional indicators also capturing informal contributions. Compared to traditional assessments, the main advantages are that the measure is instantaneous, resulting in reduced time delay, and that it relies less on formal databases.

METHODOLOGY

We seek to generate a complementary perspective by applying novel computational methods and embedding them in the current impact assessment framework of public research, therefore seizing the widely agreed potential of those applications by adapting them to our specific purpose. We capture, identify and verify the existence and extent of knowledge transfer to the economic surrounding of the DTU. To identify research outcomes, which can be attributed to the university, this study uses text-mining methods.
To implement these measures we follow systematic and distinct actions, including the sample generation, data collection, pre-processing and the application of statistical correlation measures.

Sample generation
Assuming that the private economy is an essential beneficiary of knowledge exchange, we included relevant private companies. Defined as companies

- with direct relations to DTU defined by hyperlinks on the DTU website (first-degree partners) ;
- with indirect relations to DTU defined by hyperlinks on partner websites (second-degree partners);
- with regional facilities near by the DTU (in the national context of Denmark, indicated via a Danish VTA registration).

Data collection
This study uses company websites, which consist of unstructured text-data, to identify company knowledge, products and expertise. Thus, we gathered these texts, which are available in form of online publications released by the companies themselves, and extracted them as HTML files. Associated social media entries will be included at a later stage of the project, as social media content requires specific treatment due to their special linguistic composition. The collected HTML files are pre-processed and transformed into unstructured raw text, maintaining only content and semantically relevant information. We implemented language identification parameters to extract exclusively English texts (Palmer 2010).

Text mining
To analyze the data, we apply methods from the field of natural language processing (NLP), as it provides tools for simple and advanced text analytical procedures. Text mining requires text corpora containing the relevant text fragments in form of tokens. In our case, we developed one text corpora derived from the raw text files of the company websites and a second 'reference' corpus containing an extensive sample of research publications. The university online publication database ORBIT provided the texts for the reference corpus, as this database comprises almost all publications including patents, projects, etc. made by DTU employees[1].

Pattern recognition algorithms and machine learning methods provide in-depth comparisons between the reference and the company corpus (Bird, Klein & Loper 2009). To extrapolate the important patterns, including correlations, semantic compositions and outlier comparison,

[1] http://orbit.dtu.dk/en/about.html

this study uses various available text mining methods. These include term-based methods, phrase-based methods, etc., which provide a variety of statistical tools to analyze the texts and to achieve our objectives. The analysis includes statistical measures that identify document relatedness, correlations or different types of regression parameters. Hereby, we quantify the extent of correlations between documents of the two corpora and the corpora themselves.
To detect the similarities between texts from the two corpora we use specially adapted machine learning algorithms, which extract key features from the reference corpus and compare them with the company corpus. We aim to include semantically correlated and content related approaches, to ensure the methods capture not only obvious semantic, but also content correlations. Accordingly, this approach allows us to detect shared contents among documents and enables the tracing of knowledge, which provides evidence-based insights in the 'relatedness' between the corpora.
We use statistical models, which include, but are not limited to, methods for dimensionality reduction like latent semantic analysis (LSA) (Landauer, Foltz & Laham 1998) and, for uncovering the underlying structures of the documents, probabilistic topic models for instance latent dirichlet allocation (LDA) and correlated topic models (CTM) (Blei, Ng & Jordan 2003). However, as NLP is a comparatively young field its methods undergo continuous development, therefore specific adjustments to its and models are inevitable.

Evaluation of the method
Given the identification of the extent of knowledge transfer by tracing linguistic and semantic content, we seek to extrapolate the research areas, which spread most knowledge and the companies, which make most use of university research within their proximity. To evaluate the relevance of our findings and to conclude whether our findings truly increase the understanding and measurement of (indirect) impacts we will compare our results to those of conventional measures.

POLICY RELEVANCE AND POTENTIAL

The study provides a supplementary perspective for the detection of research dissemination and impact of university innovations. Our intention is to contribute to the understanding of university performance by enhancing the detection of impacts of publicly funded research. Current computational methods provide novel possibilities for measurements allowing additional benchmarking as foundation for decision-making processes.

The goal is to provide policy makers with additional insights on the applications of university knowledge, allowing them to evaluate the benefits of government funding of research in a more holistic manner by including so far undetected, but essential impacts. This study can shed new light on the contributions universities make to economy and society.
Advantages of this novel approach are firstly, the availability of data, contrary to conventional assessments, which rely highly on university databases, which vary in quality and accessibility. Secondly, the potential to apply these measures in different regional, societal and economic contexts. Thirdly, the instantaneous nature of the measurement could capture the outcomes and the status quo almost in real-time.
After an in-depth evaluation of our approach against existing measures we will be able to verify the extent of additional information that can be drawn from this new approach. Ideally, it will provide a greater overview about (informal) knowledge exchange from universities to companies, providing a more detailed picture for future oriented decision-making.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol: O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Broström, A. (2012). Firms' rationales for interaction with research universities and the principles for public co-funding. *The Journal of Technology Transfer, 37*(3), 313-329.

Drucker, J., & Goldstein, H. (2007). Assessing the regional economic development impacts of universities: a review of current approaches. *International regional science review, 30*(1), 20-46.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17).

Huggins, R., & Johnston, A. (2009). The economic and innovation contribution of universities: a regional perspective. *Environment and Planning C: Government and Policy, 27*(6), 1088-1106.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284.

Palmer, D. D. (2010). *Handbook of natural language processing (second edition)*. Boca Raton: CRC Press.

Jongbloed, B., Enders, J., & Salerno, C. (2008). Higher education and its communities: Interconnections, interdependencies and a research agenda. *Higher education, 56*(3), 303-324.

Ramos-Vielba, I., Fernández-Esquinas, M., & Espinosa-de-los-Monteros, E. (2009). Measuring university–industry collaboration in a regional innovation system. *Scientometrics*, *84*(3), 649-667.

Vincett, P. S. (2010). The economic impacts of academic spin-off companies, and their implications for public policy. *Research Policy, 39*(6), 736-747.

B CONFERENCE: DRUID NEW YORK 2017

Title: *"Search for Knowledge: Text Mining for Examination of University-Industry Knowledge Transfer"*
Author: Woltmann, Sabrina and Alkærsig, Lars
Conference: DRUID 2017
Year: *2017*

Search for Knowledge: Text Mining for Examination of University-Industry Knowledge Transfer

1 INTRODUCTION

Decades ago a change perception of policy makers and public took place adding a new objective to universities traditional tasks, the third mission. After teaching and research, this mission is an interrelated mission, which aims to translate the teaching and research efforts of universities into economic contributions. Therefore the universities implemented various forms of knowledge transfer activities (Gulbrandsen & Slipersaeter 2007). As with teaching and research, the strategies and activities through which universities pursue the third mission vary from university to university (Ankrah & Al-tabbaa 2015, D'Este & Patel 2007). The knowledge transfer activities are highly depending on exogenous factors, such as legal frameworks (Geuna & Rossi 2011), public research policies, funding incentives (Munari et al. 2016) and the research fields the university is active in (Bekkers & Bodas Freitas 2008).

However, the key objective remains the same: the active dissemination of novel knowledge and technologies to the socioeconomic environment of the university (Zawdie 2010). From a policy perspective, the main incentive is to ensure that knowledge is actually received by the industry or other relevant third parties, which in turn are able to utilize and potentially further develop the novel knowledge (Agrawal & Henderson 2002).

In times where leading edge technologies decide about the economic development of sectors, regions and nations and about their competitive positioning in the global economy, knowledge is the key driver to foster (socio) economic development. University as research institutions have been contributing for centuries to regional knowledge development. Hence, their research plays an important role in terms of technological innovation and knowledge

creation. Research driven innovations lead to economic growth, development and increase competitiveness (Huggins et al. 2008, Vincett 2010).

Most concepts of university research impact assessment are based on the notion of knowledge transfer and is often assessed on a case-by-case basis or via somewhat limiting proxy indicators. This implies challenges to generalize the empirical findings, quantify the knowledge transfer and to draw concrete conclusions about the actual contributions to socioeconomic development. The emphasis of knowledge transfer detection lays on transfers leading to commercially relevant research driven innovations. However, even commercialized knowledge is only detected, if it is protected by patents, declared via scientific co-publications, subject to entrepreneurial activities or sold/licensed to the industry. Mostly it needs to be generating some sort of direct revenue. It is acknowledged that commercialization is only one aspect of knowledge transfer and that several other levels are existing, including the creation, sharing and implementation (Sung & Gibson 2000). Moreover, commercialized inventions are estimated to represent only a small fraction of the actual transferred and used knowledge(Agrawal & Henderson 2002, Drucker & Goldstein 2007). But the transfer of not commercialized knowledge is even harder to trace and measure, which leaves the research community, up until now, with lack of understanding of its occurrences and very limited metrics to capture these. Therefore, the actual knowledge transfer often remains an approximation.

Various attempts to identify and quantify knowledge transfer on diverse levels show that current scholarly literature fails to provide metrics to capture the complete knowledge transferred from universities to the industry (Malerba 2007). Different proxy- indicators and assumptions about knowledge transfers, spillovers and their channels are employed to compensate this lack of direct measurability (Cheah 2016, Lin 2016, Salter & Martin 2001). These indicators are not holistic, but as F. Malerba states for one of them: *'the use of patent citations in order to examine knowledge flows and networks is a very fruitful research direction, provided that one is aware of their limitations and uses them jointly with other qualitative and quantitative indicators'*(2007, p. 13).

Given these evident research challenges, the three key objectives for this study are i) to identify novel methods that allow direct identification of common knowledge contents between universities and industry using additional data sources, ii) to identify whether these common contents originate from the university and iii) to capture common areas of knowledge in the geographical proximity of the university. The overall goal is to enable a flawless detection of research knowledge contents ensuring generalizable and comparability of findings regardless of the case. We propose a novel method that

identifies and to a certain extent quantifies knowledge transfers from universities to the industry. We aim to capture the transfer without focusing on channels, commercialization or transfer mechanisms.

We are proposing contemporary computational methods from the field of natural language processing (NLP) including text mining and statistical learning tools to adapt a novel metric to measure knowledge transfer. Using text material from corporate websites and academic publications, we aim to identify common and related topics. We use pattern recognition tools and similarity measures to identify overlapping and coherent contents. Point of departure are the scientific texts, assuming that publications are limited to developers of an invention or novel research insights. Hence, we derive content from various scientific texts by organizing them into text corpora and extracting their key concepts. Afterwards we identify similarities to the commercial websites.

The clear identification of common knowledge areas of a university and its economic environment, the traceability of shared concepts, knowledge and technologies provides additional tools to enable scholars, policy makers and practitioners to perform more in-depth analysis. The flexibility of the tools and their potential for adaptation make them useful in various contexts.

2 THEORY

The body of literature on knowledge transfer between universities and the industry contains several interconnected topics. These focus on issues including firm and university characteristics when engaging in collaborative activities (Ankrah & Al-tabbaa 2015, Brostroem 2012), the identification and verification of knowledge transfer channels (Agrawal 2001, Grimpe & Hussinger 2013, Schartinger et al. 2002), policy implications, funding and legislative regulations for universities engagement (Munari et al. 2016), the role of academic fields or industry sectors (Bekkers & Bodas Freitas 2008, D'Este & Patel 2007), the impact of university research including economic, societal and political dimensions (Drucker & Goldstein 2007, Jong et al. 2014). Key aspects are, among others, the measurable identification of knowledge transfer itself and its subsequent commercialization successes (Thursby & Thursby 2002). This diversity shows that the understanding of university-industry collaboration and subsequent knowledge transfers including its impacts are highly investigated by now. This reflects a well-developed academic research area, which led already to an advanced policy understanding and elaborated empirical studies (Munari et al. 2016)

Given this broad understanding it is evident that knowledge development

and knowledge transfer are highly interrelated and increasingly relevant topics for academic research. Today the area is a multidisciplinary empirical research field including well developed literature on knowledge transfer and university-industry collaboration.

2.1 Knowledge

The majority of studies, however, lack a clear definition of the concept of 'knowledge', as well as of 'knowledge transfer'(Liyanage et al. 2009). The concepts seem to be commonly agreed and the framework seems widely understood; however we see the need to use definite concepts for this paper.

The term itself is highly debated and conceptualized in various philosophical approaches; to limit it to a reasonable scope, we focus on the definition relevant for this particular context. Therefore we use the foundations of knowledge management theories. Knowledge management focuses on two main types of knowledge: explicit and tacit knowledge. *"Explicit or codified knowledge involves know-how that is transmittable in formal, systematic language and does not require direct experience of the knowledge that is being acquired (. . .)"* (Howells 2002, p. 872). Tacit knowledge, on the other hand, is described as *"non-verbalised, intuitive and unarticulated knowledge"* (Polanyi 1962). It represents a form of know-how that is developed by informal acquired behaviours and procedures (Howells 2002, p. 872). For the purposes of this study, we refer to knowledge solely in terms of the concept of explicit knowledge.

Furthermore, this study focuses only on research related and novel knowledge. The scope comprises knowledge and technologies, which are potentially relevant for future innovation processes and are novel to the scientific community. This includes all recent research outcomes by a university, but excludes widely known and commonly accepted knowledge. Therefore, alone novel scientific insights, technological innovations, like leading edge technologies shape the scope of this study.

2.2 Knowledge Transfer

Knowledge transfer and technology transfer are in the body of literature extremely interrelated concepts and thus often used in an exchangeable manner (Agrawal 2001, Grimpe & Hussinger 2013, Sung & Gibson 2000). We, however, focus on knowledge transfer overall, but acknowledge that the term 'technology transfer' is, in certain cases, a more accurate description of the issue. A closer look at the literature on knowledge and technology transfer reveals that most studies omit to deliver a clear definition of knowledge

transfer (Liyanage et al. 2009).

We aim to set common ground by explicitly defining knowledge transfer, in accordance with Argote and Ingram (2000, p. 152), who define knowledge transfer broadly as : *"the process through which one unit (e.g., individual, group, or division) is affected by the experience of another."* However, this study requires an additional definition, which includes more precisely the mechanisms and outcomes of the transfer. Hence, we expand this notion by including the aspect from the notion of Liyanage et al. (2009, p. 123) describing that *"(. . .) a knowledge transfer process has two main components, i.e. the source or sender that shares the knowledge, and the receiver who acquires the knowledge."* For our purpose it is crucial that the emphasis lays on the fact that the for the transfer to be evident it must be measurable on the receiver's end. While the variety of definitions in literature, given the above mentioned constrains, this paper will follow closely the final definition again suggested by Liyanage et al.(2009, p. 122) who sees knowledge transfer as *"(. . .) the conveyance of knowledge from one place, person or ownership to another. Successful knowledge transfer means that transfer results in successful creation and application of knowledge in organizations."* This definition is particularly appropriate as it includes the necessity of utilization of the transferred knowledge, pointing out the criteria for successful transfer.

2.3 Formal and Informal Knowledge Transfer

The research on university based knowledge transfer to industry is divided into two main categories: "formal" and "informal" knowledge transfer. Some scholars define formal knowledge transfer mechanisms as such, which eventually *"result in a legal instrumentality such as, for example, a patent, license or royalty agreement (. . .)"* (Arundel & Bordoy 2008, p. 642), while informal knowledge transfer is seen as a transfer resulting from different forms of informal communication, including consulting or collaborative research (Link et al. 2007). However, we view this definition of 'informal' transfer as still defining mainly a formal forms of transfer, as it is still based on formalized agreements pursued under contracts between the two entities. Hence, research joint ventures, and university-based start-ups would be a form of formal knowledge transfer (Link et al. 2007). Therefore we follow a less common understanding of informal knowledge transfer including transfer, which is not based on property rights and the exchange may refer to personal contacts, informal use of data bases, workshops or similar. Here the obligations between the partners are more normative than actually legal (Fernández-esquinas et al. 2015, Grimpe & Hussinger 2013, Link et al. 2007). Overall, it is evident that the main attention in university knowledge transfer has been

given to the formal knowledge transfer including their mechanisms, successful commercialization of inventions, impacts and similar (Link et al. 2007). We aim to consider both types in our study.

2.4 Qualitative and Quantitative Approaches

Informal aspects of knowledge transfer are mostly studied in qualitative studies, like case studies. These are often capable of capturing various transfer channels, motives behind collaboration and similar. They provide in-depth insights into potential transfer mechanisms, motivations and rationales of the actors and more (Brostroem 2012, Franco & Haase 2015). Qualitative studies focus often on single cases, including at best national contexts and provide in-depth understanding about potential impacts and benefits of certain activities, projects or particular innovations (Ankrah et al. 2013). Qualitative case studies fail to provide measurable and generalizable results and offer in-depth insights only into very specific scenarios. This limits the comparability of the findings, in particular since knowledge transfer is highly depended on exogenous influences like policies, legal structures, funding etc. Moreover, many studies fail to actually verify the content of the transferred knowledge, or the extend of it (Rothaermel et al. 2007, Salter & Martin 2001).

Quantitative studies, on the other hand, often deal with the overall contributions of formal knowledge transfer between universities and industry. They aim to capture mainly economic and/or socioeconomic impacts. The level of analysis ranges from firm and university to regional and national comparative studies. However, these studies mainly capture commercialization of products or technologies and revenue generating usage of more or less finalized inventions derived from research knowledge (Cheah 2016). Main proxy indicators for quantitative studies are, among others, licenses and license agreements (Jensen et al. 2003), patents, including patent citations (Arundel & Bordoy 2008, Thursby & Thursby 2002), co-publications by firm and universities (Tijssen et al. 2009), and different kinds of entrepreneurial efforts, like university spin-outs and their generated revenues (Vincett 2010).

However, these indicators face long-standing criticism about their incapability to capture the majority of transferred knowledge. Some of these proxies simplify the transfer to a plain commercially measureable value (e.g. royalties) and others fail to capture the collaborative relationship and focus on potentially never utilized knowledge (e.g. co-publications) (Cheah 2016, Lundberg et al. 2006). Many quantitative studies combine the investigation of formal knowledge transfer, in terms of commercialization with qualitative methods, like expert interviews, to capture a more holistic picture of the knowledge transfer and the collaboration in general (Cohen et al. 2002, Siegel

et al. 2004). The indicators for economic impact of quantitative assessments are *'(. . .) difficult to obtain and generally suffer from long lag times between public investment and outcomes'* (Arundel & Bordoy 2008, p.6). Besides, it has been pointed out that they fail to provide a holistic picture. Some scholars argue that the measurements are only accounting for very low percentages of actual knowledge transfer (Cheah 2016, Lundberg et al. 2006).

Given the limitations of contemporary empirical work we contribute to the body of academic literature by addressing some of the deficits of the current metrics. We aim to provide a novel measurement method that helps to diminish the limitations. Our method provides in-depth insights about the transferred knowledge. Ideally even traceable to university department, or academic scientist level. It is supposed to provide statistical correlation measures comparable to the ones of patents or licenses analyses. We offer a great extend of independence from the concepts of formal and informal knowledge transfer channels, aiming at an additional more holistic way of capturing knowledge transfer. The approach is less dependent on the examination of external and internal circumstances, as it measures the transferred items and does not require additional assumptions about the potential of knowledge exchange. However, the commercialization is not directly measurable in terms of patents or revenues , but in combination with the traditional measures could provide comparatively precise estimations.

3 METHODS

Texts contain information, and extracting this information has become an increasingly developed part of today's research fields of machine learning and computational linguistics. Enormous insights in various disciplines were generated via the use of text mining tools over the past decades. Content and sentiment analysis are of increasing relevance in computer science and machine learning during the past decades and the tools advancement is getting more and more promising (Chapman & Hall/CRC 2010, Collobert et al. 2011).

In our case text mining is an appropriate strategy, since text material can be a sufficient data source to detect knowledge transfer. First, academic publications in form of scientific texts, such as journal articles, conference proceedings or books, contain the main outcomes of scientific research. They are seen as output and dissemination channel of university research (Stahl et al. 1988, Toutkoushian et al. 2003). Therefore these publications are texts containing data for all major research findings of a university.

Second, online presences (like websites) are media for companies to dis-

play their novel products, services, R&D strategies and innovations, as websites, blogs, videos or social media entries. These online presences of firms are mainly in text form and firms place high value on these to ensure their visibility for potential consumers and investors leading to regular updates and R&D descriptions (Branstetter 2006, Heinze & Hu 2006).

These two types of texts provide insights into the use and generation of knowledge. Therefore, the use of statistical tools from NLP is an ideal approach to identify commonalities in terms of correlations.

3.1 Pre-processing

Text pre-processing converts unstructured raw text into statistical and computational useful units. Pre-processing is part of any text analytic procedure and might very well be decisive for the later outcome of the analysis. The quality of the results is highly depending on the thoroughness of the pre-processing. The main objective is to capture all relevant characters, erase obsolete items. To identify actual words via the detection of word separation (tokenization). This enables the application of further text mining methods (Paukkeri & Honkela 2010).

Pre-processing of text includes:

- To define word boundaries as white space,
- To delete unwanted elements (e.g. special characters, punctuation and numbers),
- To convert upper case to lower case characters,
- To remove 'stopwords'[1],
- To stem the words [2].

Results of our pre-processing revealed some challenges in the case of the academic abstracts. These abstracts contain, for instance, chemical formulas and notations, which rely heavily on numbers and/or special characters. These are unfortunately lost during the course of the pre-processing. The only possibility to later identify the same formulas and to use them for similarity measures is the assumption that the removal will always result in an

[1]Stopwords are the most common words in a language, which are not carrying content relevant information

[2]Describes the process of reducing words to their word stem or root form. It is a process for removing the morphological endings from words: connected, connection, connections become 'connect'.

identical character string. The result may not be identifiable as the specific formula, but still provide a match [3]. However, particularly short strings derived, from formulas or notations, are lost during the pre-processing. Some terms seem to be the result of poor pre-processing, but are in reality just a representation of specific models, formulas or project names shrunk to a unidentifiable string of characters. The pre-processed texts are merged into structured units: the text corpora. All the following methods are based on these corpora, which are an a way of structuring documents as well as organizing them into meaningful content related units.

3.2 Document Term Matrix:

A document-term matrix is the most common vector space representation of a document corpus. It contains the feature (term) frequencies for each document. Rows correspond to documents and columns to terms.

A document-term matrix is usually generated from pre-processed corpora, which results in a representation of semantically and contextual relevant terms (Chapman & Hall/CRC 2010). As document-term matrices are usually highly dimensional and sparse; hence many of the current models aim for sensible dimensionality reduction (Berry & Castellanos 2007). In a document-term matrix the element at (m,n) is the word count (frequency) of the i'th word (w) in the j'th document (d).

$$\text{Document-Term Matrix}(w,d) = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

Various schemes determining the value of each entry in the matrix can take have been developed, which are the term weighting schemes. The weight for each term can be derived in various forms from frequencies of the term occurrences.The weighting of terms differ widely depending on the models used. Common weighting schemes include, among others:

- The binary weighting, the entry takes values 1 or 0 depending on whether or not a term occurs,
- Term-frequency (TF), the actual number of times a term occurs,

[3] In some cases HTML tags prevent the identical construction. In this case we did not find a way to identify the matching strings.

- Term-frequency, inverse document frequency (TFIDF), uses TF but assigns higher weight to terms that occur only in a small number of documents.

3.3 Term-frequency, inverse document frequency

For the purpose of this study we chose to use the TFIDF indexing to determine the 50 most characteristic words per document. In case of the academic abstracts we retrieved often less than 50 words. The reduced dimensionality enabled a comparison of keyword lists with each other. We use these lists, generated for each document, to identify common terms between two types of documents, abstracts and website pages.

The TFIDF is a simple numerical indexing method, which has been applied in various contexts (Franceschini et al. 2016, Zhang et al. 2016). It has proven to give respectable results on its own, especially considering its simplicity. However, it is also used in various more advanced models, such as Vector Space Model (VSM) or Latent Semantic Analysis (LSA) (Mao & Chu 2007).

TFIDF can be used to enable a dimensionality reduction providing a small set of content relevant terms, which capture the main content of a document. Further, TFIDF is an indexing scheme that allows identifying the most relevant words by extracting the words most unique to a given text. The principal assumptions are simple: a word that occurs often in a document is relevant for its content (of course after the stopword removal), but words that are additionally used in many documents are less specific for the single document and therefore less relevant. Hence, frequent words that are used in many different texts are seen as carrying less contextual information and obtain a lower score in the TFIDF weighting. The scheme has different proposed calculations, but most commonly the TFIDF weight is calculated by multiplying the term frequency TF, the number of times word w appears in document d; and the inverse document frequency IDF, which is the logarithm of the total number of documents D divided by the number of documents that contain the word w denote dw.

$$TF(w,d) = \sum w_i$$

$$IDF(w,D) = log(\frac{D}{dw})$$

$$TFIDF = tf(w,d) \times idf(w,D)$$

The TFIDF approach suffers from three main shortcomings:

First, as it calculates term weight based on term frequencies, it might represent only the content of a text fragment, since terms with a high frequency may be only used in a certain part of a document. This is a major draw back especially for long texts

Second, IDF assumes that terms, which rarely occur over a collection of documents, are more content related, while in reality it just makes it more distinct from the other documents in the collection. So a corpus about, for instance, water issues would probably score the term 'water' low, which does not capture the reality of the document content.

Third, empty terms and function terms, like adverbs or modal particles, are often assigned too high scores, which leads to inaccurate weight. Unfortunately, even a thorough stopword removal is not preventing this from happening (Xia & Chai 2011).

3.4 Latent Dirichlet Allocation

LDA is an application of topic modeling and is a fully automated method based on statistical learning, which aims to identify latent (unobservable) topical structure in a text corpus (Blei et al. 2003, Griffiths & Steyvers 2004). LDA extracts underlying structures of texts and translates them into topics, which are composed of terms that are assigned together with a certain probability to each topic.

LDA works as follows, described by Grün and Hornik (2011, p. 4) and Ponweiser(2012, p.15):

1. For each topic: decide what words are likely (term distribution described as $\beta \sim Dirichlet(\delta)$
2. For each document:
   (a) decide what proportions of topics should be in the document, (topic proportions defined by $\theta \sim Dirichlet(\alpha)$.
      i. for each word in the document:
         A. choose a topic ($z_i \sim Multinomial(\theta)$).
         B. given this topic, choose a likely word (generated in step 1.) from a multinomial probability distribution conditioned on the topic $z_i : p(w_i|z_i, \beta)$.

To select the optimal number of topics (K), we chose to approximate the marginal corpus likelihood (depending on K) by taking the harmonic mean of the corpora after applying the LDA. The harmonic mean takes one

chain of samples as argument to first collect all sample log-likelihoods and subsequently calculates the harmonic mean of these likelihoods. This is an approximation of $p(w|K)$, i.e., the likelihood of the corpus given the number of topics (Ponweiser 2012).

We limited the maximum number of topics for the corpora of websites in each case to 50 and found that only 3 web corpora might have benefited a larger number, we chose this limitation for computational efficiency reasons. For the academic abstracts, the calculation for the optimal number of K was set to a maximum of 200. This was chosen due to the diversity in academic specialized fields. However, results show that the optimal number of topics only rarely exceeded 50.

We chose to set the hyper-parameter (α and β) so that they allow a more diverse topic distribution over a single document by enforcing more topics per documents with lower probabilities[4]. This is appropriate, since we are not trying to classify the documents but working to fine grain the content of the documents to an extent that captures context and topics of text snippets. To improve the performance we added one pre-processing step that excluded terms, which occur in more than 90% of the documents in the document-term matrix. The resulting topics are very specified, especially after the additional pre-processing step.

We used the obtained 50 words per topic with the highest probability for this particular topic and returned them as list of keywords. We compared to other lists of topic keywords from LDAs from academic corpora and web corpora. The resulting topic pairs show the most similar corpora in terms of their underlying structures.

3.5 Jaccard Similarity Coefficient

For the similarity measure between the sets of identified keywords found by applying TFIDF or LDA, we used the Jaccard similarity coefficient as metric. It is a statistic used for measuring the similarity between sets. The Jaccard similarity is based on the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, 1 indicating most similarity (identical sets) and 0 indicating least similar: no common feature in the two sets. Given the set of keywords from one document of the publication database denoted K_A and the second set of keywords from one page of the websites denoted K_B, the Jaccard similarity denoted $J(K_A, K_B)$ is obtained with:

[4]We used Gibbs sampling in the LDA model to draw from the posterior distribution. For more information on determining the posterior probability of the latent variable, refer to (Grün & Hornik 2011)

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A|+|K_B|-|K_A \cap K_B|}$$

We chose this similarity measure as it only includes element presence in a given set. It is applicable for the LDA and TFIDF generated keywords, as it does not need an input of scores or probabilities, which would not be comparable since they resulted from different corpora. Another advantage is the low computational expense, making it attractive for a basic similarity assessment. However, this could of course be refined by applying additional similarity measures to find more accurate matches. The thresholds for a minimum similarity for further examination were chosen based on previous manual examination; meaning that we would only consider keyword lists with a certain minimum Jaccard similarity as relevant for the manual inspection and potential text matching. However,the Jaccard similarity tends to benefit smaller sets. Hence, we decided to set a common threshold to a minimum of 0.13 and another used indicator threshold consisted in multiplying the Jaccard index with the intersection of the two sets, giving higher weight to sets with a higher amount of common words. A set of pairs with Jaccard Similarity lower than 0.15 needs more than 7 words in common in order to pass the criteria, while set pairs with Jaccard Index higher than 0.15 can have smaller intersections.

3.6 Sample

The following description of the sample is divided into the generation of the text corpora, representing a) university and b) industry knowledge.Two main data sources were needed for the analysis: academic publications representing university research output and a collection of relevant texts from firms.The methods are applied to data from the Technical University of Denmark (DTU).

The university publication database, Orbit, provided data that include a collection of academic abstracts from university research publications. These abstracts present a summary of the main research outputs by employees of the university between the years 2005 and 2016.

Firm websites are the second data source for this study, providing the company knowledge. Criteria for relevant websites are a) a national (Danish) registered branch of the firm b) at least some English fragments of the firm website, and c) the firm must have been a 'partner'[5] of the university between 2013 and 2016.

[5]The types of relevant partnerships are explained in the next section of the article.

Publication database
Given legal challenges to obtain a comprehensive sample of full text publications, we chose available abstracts to serve as proxy of the university's research output. Approximately 60% of the publications during those years were stored with an abstract. The data availability in the database increases from 2012 onwards here approximately 80% of all abstracts are publications are available. The selection criteria required that an entry is a) published between 2005 and 2016, b) has an available abstract and c) is least co-authored by one member of the university staff. These criteria resulted in 43,745 hits while the total number of all publication records is 76,627.

We classified the abstract by their assigned departmental codes, provided by the database, to enable a pre-classification of the texts on the basis of their research field. Both methods, LDA and TFIDF, perform better on more content coherent corpora, as this to enables a better performance of the statistical analysis. In particular in the case of the LDA, one single corpus as input would result in identifying the overall research giving us almost no additional insights. The classification resulted in 24 separate research fields, while three of these were irrelevant for the academic output of the university [6]. The collection of corpora based on academic abstracts is in the following referred to as 'academic' corpora or by the individual research field, if it is relevant for the interpretation of the results.

Firm Websites
To identify relevant firms for the sample we performed two major steps. First we collected based on Danish companies with a formal connection to the university, namely a collaboration contract. We identified 686 Danish firms, which had a contract with the university between the years 2013 and beginning of 2016. The firms in this sub-set operate mainly in technology intensive sectors and are firms with strong R& D divisions. Therefore it included companies with contents similar to the research performed at the university. Second we generated a network on the basis of hyperlinks between websites using the university as point of origin, identifying the university's partners linked to the university website. Partners of those partners (second-degree partners) of the university were hereby also identified and added. These websites content were downloaded and stored as HTML files. The list of examined websites contained many online service platforms, including for

[6] We excluded i) publications registered to the university administration, ii) publications registered to the bachelor program, and iii) one set that was directly linked to a large firm (this could have biased the findings significantly as the firm is directly involved in several hundreds of dedicated publications).

example public transportation sites, yellow pages and firm registries. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample. The online text samples were collected between August 2016 and November 2016.

To ensure a connection of the firms to Denmark each page of each website was subsequently scanned for a Danish firm registration number (CVR) and in case one was found the website was added to the sample.Unfortunately, in Denmark, universities, schools for higher education and other public entities are registered via the firm registration number; they had to be manually excluded. Finally the language of the website and/ or the sub-pages was verified and only content with more than 60% English content was stored.

The total sample contains 599 Danish websites, containing English pages with a total of 148939 sub-pages (documents). 464 websites provided more than 5 English pages and were converted into single corpora and used for the TFIDF application. Due to more extensive pre-processing procedures the number of useful websites for the LDA was 404 websites. The number of pages and length of documents varies a great deal between the firm websites. Some provide just an English summary for their main contents, while others, often multinational firms, have their entire website in English, which influences the model performance. One major drawback in our sample collection is the partial absence of PDFs or similar formats stored since these require special treatment for each format.

4 RESULTS and DISCUSSION

The results of this study are divided into the application of the different methods. We aim to provide in-depth details about the performance of each single tool and algorithm. Additionally, we describe interrelated components and the results generated via a combination of those methods.

4.1 LDA

As described earlier, the LDA is a representation of the hidden structures of the content of a given text corpus determined trough a set of topics. The main words per topic show an adequate representation of the overall topics of the corpora. It means that for example the themes of the abstracts in the corpus of Chemistry are represented in 37 topics.

This extract shows that the words are representing overall topics of the academic corpus quite satisfactorily.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	enzym	bind	dynam	forc	oligosaccharid	indic
2	domain	site	vibrat	particl	branch	chain
3	amino	conform	motion	hydrophob	carbohydr	complet
4	residu	enzym	coupl	friction	donor	size
5	express	residu	excit	layer	polysaccharid	mean
	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	speci	chemistri	optim	raman	treatment	situ
2	ester	analyt	factor	spectra	strontium	redox
3	previous	research	occur	band	bone	electrochem
4	iridoid	uniqu	design	laser	reduc	stm
5	isol	european	reveal	mixtur	treat	microscopi

Table 1: The top 5 words for academic topics

However, in the case of the websites we observe a more diverse outcome with less diversification among the topics. One could say that the single topics within the web corpora are less coherent and provided more heterogeneous themes than the academic ones. The keywords of the topics of the web corpora seem to be more generic. This is attributed to the length of texts (abstracts are shorter than websites) and the content diversity (abstracts contain mainly one single theme).

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	engin	shipment	museum	binocular	environment	spectacl
2	graduat	ship	exhibit	wetzlar	wast	optician
3	knowledg	mention	west	riflescop	conserv	coat
4	opportun	nation	fascin	mechan	water	distanc
5	student	mobil	photon	assembl	bird	wearer
	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	altern	cataract	hunt	eur	lamp	outstand
2	attach	iol	binocular	carlzeissstraß	slit	packag
3	booth	iolmast	outdoor	carlzeisspromenad	oct	rock
4	confid	biometri	spot	auxiliari	cirrus	broad
5	frequent	refract	passion	consolid	fundus	complex

Table 2: The top 5 words for website topics

In order to capture relevant pairs of academic abstracts and website texts we decided to combine three different approaches to compare the keywords between the LDAs.

Academic topics vs. Websites topics

The first approach is to compare the similarity of each topic from the web corpora with each topic of the academic corpora. Each topic identified by the LDA is converted into a simple word list. The Jaccard similarity measure was used to compare and match word lists. The number of list comparisons made to assess the topic similarity was 17,417,025.

Given the chosen Jaccard threshold (as explained in the method section), only $1.3 \times 10^{-4}\%$ (23 pairs) exceeded the required similarity for the topic per topic comparison. 9 of these matches were Danish language fragments embedded in the topics. They were removed and the remaining 14 matches were manually inspected. The common words are comparatively generic, but show relatedness. The common words indicate a clear overlap in the topics, but are slightly unspecific, as they lack any reference to relevant models, notations, formulas or relevant proper nouns. This might very well be a result of the fine tuning of the LDA and the LDA's potential to identify fields rather then specific content.

	12 common words between 2 topics word list
1	function
2	gene
3	dna
4	express
5	isol
6	microorgan
7	cell
8	strain
9	bacteri
10	bacteria
11	communiti
12	popul

Table 3: The top 10 most common words academic topics vs website topics

Interestingly most common words within the remaining 14 pairs were based on 8 distinct corpora. One corpus alone accounted for almost a third the total matches. The corpus that accounted for that many pairs was the corpus, which contains 'diverse' research areas that could not meaningfully be fitted in any other department. This was surprising, but given the corpus diversity it would represent the most mixed research topics which are present in several websites.

This corpus lacks the coherence and specificity of the other academic corpora and is therefore highly

Academic topics vs. Websites

The second approach is to combine all topics of one website corpus into a word list representing the entire corpus and compare this web corpus list with each individual topic of the academic corpora. The academic topics are more defined or specific and are therefore the adequate choice. The goal is to find new relevant abstracts and website matches.

Given a new Jaccard threshold (0.45×10 words), due to the increase in number of potential common words we obtained $4.6 \times 10^{-3}\%$ (17 pairs) matching pairs for 370.575 comparisons. No purely language based pairing occurred in this instance.

The most common words were comparatively generic and the number of their occurrence was rather low, with a maximum of nine co-occurrences, showing that the pairing was based on comparatively diverse keywords.

	10 common words
9	aim
6	focus
6	requir
6	year
6	dtu
6	challeng
6	tool
6	recent
5	continu
5	research

Table 4: The top 10 most common words topics vs websites

Departments vs. Websites

The third approach is to create one combined word list for each of the web corpora and a second combined word list for each of the academic corpora. These comparatively long keyword lists are subsequently compared with each other.

The average keywords list per academic corpus contained 1900 keywords, of course depending on number of topics. The web corpora had an average 2300 keywords per corpus list.

We compared the total keyword lists of both and find the next pairs. For the corpus based keyword comparison, we had only 8505 comparisons and due to the high number of words per corpus list we had to adjust the Jaccard similarity thresholds to a minimum Jaccard similarity score of 0.2 and a minimum word intersection matching of 0.18×500 words. We obtained 41 positive matches, which contained non based on foreign language fragments. The main corpus matches were based on several Departments. The combination of websites was, again, diverse without any clear patterns or specific corpus domination.

We compared in the outcome of the LDA comparisons with the manual investigation conducted after the application of the TFIDF.

	Academic Topic vs. Web-pages Topic	Academic Topic vs. Websites	Departments vs. Websites
1	Diverse[7]	Energy Conversion	Management Engineering[8]
2	Mechanical Engineering[9]	Electrical Engineering	Diverse
3		Diverse	Civil Engineering
4		Civil Engineering	Compute Math

Table 5: Results for the LDA

4.2 TFIDF

The TFIDF indexing resulted into a set of keywords for each single document, for both academic and website corpora. This resulted in 3,343,890,411 comparisons. Every match (a comparison, which exceeds the threshold) was was stored as text pair for later manual assessment. We excluded multiple matches between the same academic abstract and the same website (but different web-page within the site). We kept only the match with the highest score Jaccard similarity score, because some companies display the same texts on more than one page.However, we left matches that referred to the same university department and the same website, but to a different abstract in the sample. Since abstracts of the same department are less likely to be identical than text snippets on the same website.

We found exactly 100 pairs that exceeded the chosen Jaccard similarity threshold. However, after some manual investigation of the outcomes we found that we had to exclude matches that were based on country names [10] Additionally,some matches were based on other language fragments entailed in the abstracts and the websites. These were mainly displaying German,

[10]A full exclusion of country names for future applications is considered, but seemed not necessary for the current sample.

French and Danish content. This was solved by the application of a simple language filter which identified all Danish, German and French key words in the sets and removed the match in case it had more than 5 hits. After this removal, the total matching pairs was decreased to 88.

The left matching keyword matches were manually assessed and we found that the dominant words were highly diverse and many entailed not real words. The character strings derived from trademarks, proper nouns, models, software names and formulas, were most present and helpful to identify relevant matches. Terms like 'novirhabdovirus', or 'mxgs' (Gamma ray Sensor module (MXGS)) account for a number of hits.

After the quality assessment of the keywords per document we paired the relevant texts and manually checked their similarity. We then classified the pairs into 7 different categories.

1. Identical topic = University contribution
2. Identical topic = Potential university contribution
3. Identical topic = Unlikely university contribution
4. Identical topic = News paper article about university
5. Different topic = No match in content
6. Identical topic = University contribution to a public entity
7. Unclear = could not be classified

The manual classification was undertaken taking into consideration the full text publication, since in many cases the abstract would not provide sufficient information to establish whether contents are actually related.

Additionally, we had to make qualitative distinctions between the entities, which display the university research, since all of them fulfill our requirements, but not all of them are actually private firms using the research. We found newspaper articles presenting university research; we found several public entities (with CVR numbers), which use and promote the university research. These can be seen as correct pairing from the TFIDF (true positives), but show that the differentiation between public and private entities needs to be improved. To exclude newspaper articles and there like might be rather challenging, but with a news registry this might be achievable.

A result summary is presented in Figure 1 and Figure 2

The results present a relatively high number of content related, but unlikely truly related matches (category 3) this shows that the TFIDF finds

related content, but an additional measure to minimize these hits would be beneficial. Even more so for the pairs off classification label 5, which provided not even the same content and result in a high number of false positives.

Figure 1: Results classification

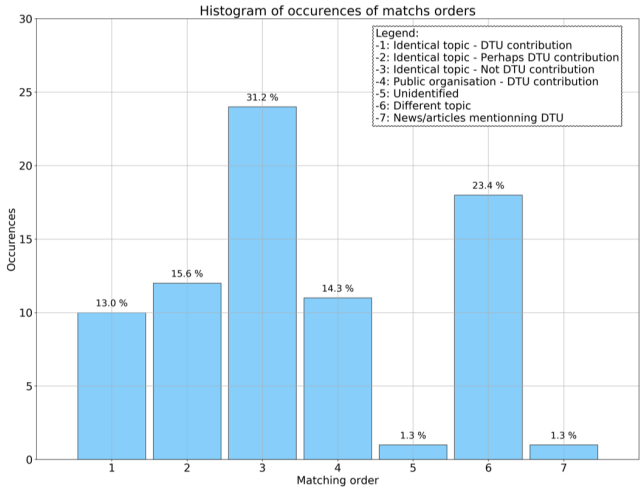
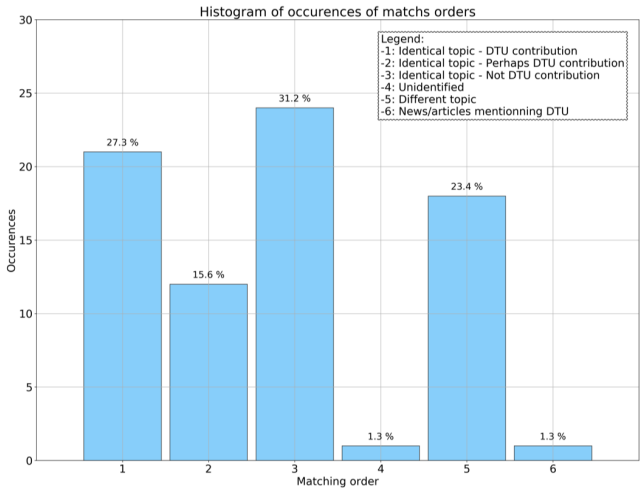


Figure 2: Results classification with classes 1, 4 and 7. combined



The academic departments that occurred in the pairs of documents were diverse, but certain departments dominated the pairing. The main academic departments found to be most often in the pairs were as indicated in Table 6 [11].The websites on the contrary were very diverse and no particular websites were matched more frequent than the others. We had a maximum of five hits to the same web-page.

	Category 1	Category:1 and 2	Category: 5
1	Bio Systems	Bio Systems	Management Engineering
2	Electrical Engineering	Electrical Engineering	Mechanical Engineering
3	National Space Institute	National Space Institute	National Food Institute[12]
4	Diverse[13]	Diverse	
5		National Veterinary Institute	
6		National Food Institute[14]	

Table 6: Departments present in the results

The overall identification of the TFIDF went surprisingly well, given that the majority of academic abstracts contain only between 100 and 200 words. In many cases the human verification needed additional information to the abstract (like the full text of a publication) to ensure the match was an actual true positive.This is very promising; especially considering the improvements that would be possible with a full text sample from the academic publications. For example the following abstract text provides enough textual information for a class 1 pair:

> *'A method for reproduction of sound, based on crosstalk cancellation using inverse filters, was implemented in the context of testing telecommunications devices. The effect of the regularization parameter, number of loudspeakers, type of background noise, and a technique to attenuate audible artifacts, were investigated. The quality of the reproduced sound was evaluated both objectively and subjectively with respect to the reference sounds, at points where telecommunications devices would be potentially placed around the head. The highest regularization value gave the best results, the performance was equally good when using eight or four loudspeakers, and the reproduction method was shown to be robust for different program materials. The proposed technique to reduce audible artifacts increased the perceived similarity.'* (Gil Corrales et al. 2015)

[11]Category entails category 7, since it is a performance indicator of the text mining application and not a performance measure for our firm/ non- firm classification.

4.3 Combining Results: LDA and TFIDF

Given the findings from the TFIDF and LDA potential improvement of the outcomes is possible if the results are combined. Hence, we used the different outcomes of the LDA to identify academic corpora that are more prone to generate false positives. The TFIDF true positives and false positives were reduced to their departments and these were compared to the LDA results.

Corpora suggested as relevant in the LDAs are mainly accounting for general content matches by the TFIDF, meaning that these corpora are at least having related contents with the matching websites.

The topic to topic comparison does not capture the most relevant corpora for the later found matches. However, this could suggest that there are more matches within the data, which are not yet uncovered. We assume that additional measures would be adequate to improve this outcome. However, we found that if a department was more than once identified in the academic topic to website topic comparisons the department was likely a true positive pair.

In the case of the comparison of academic topics to website corpora lists, we found that a department, found in this combination, is also likely to be a true positive. None of the identified departments was only in the false positive collection, hence a match here also suggests a content relation between texts.

In the final case, one keyword list for an entire website corpus was compared to one keyword list for an entire academic corpus, we found that one department was only present in this pairing and here over-represented. It also accounted for 25% of all false positives in the TFIDF application. The academic corpus is the one of the department of Management Engineering and is noticeable due to its generic keywords in both LDAd and TFIDF. This could suggest that this last LDA comparison focuses on similar patterns as the TFIDF when identifying the false positives.

If a department was not in any of the LDA comparisons, it was also likely to be a false positive. However, it was not possible to detect the false positives that share a department with a true positive. For this matter an investigation of the websites could be considered.

5 CONCLUSION

What are the implications of these results for the use of text mining as metrics in the studies of university-industry knowledge transfer?

First, our results show that there are many occurrences of commercially used knowledge transfers, which are not necessarily only identifiable via

patent, license agreements or similar. It clearly shows that university research is used and displayed on firm websites and that these instances are computational traceable. Hence, we are confident that we can confirm Agrawal and Henderson (2002) findings who stated that patents present only a small fragment of the knowledge transfer between universities and the industry. Even though our sample size is not large enough to estimate the extent of additional knowledge transfer that can be identified via our method, we can say for certain that we captured additional knowledge transfer.

Second, we see that our findings are in agreement with the notion that certain academic fields are more prone to knowledge transfers than others. This confirms the notion that the transfer of applied sciences is more frequent than the one of basic research. However, since our sample is currently limited to one case it does not yet provide generalizable results.

6 LIMITATIONS

The limitations of our study are numerous and are technical as well as conceptual. First, the data on academic research outcomes is limited, since abstracts hardly display the true output of the research. The use of abstracts was necessary due to availability issues and copyright issues for full-text publications. However, in future we aim to complement the data with full-text publications. Second, the manual classification is not ideal as it is time intensive, especially since the text pairs are often hard to understand and therefore difficult to classify. It often requires expert knowledge from the specific research field. We hope to address this shortcoming in future by building a computational classifier that would at least propose a first potential classification, which would only have to be verified by human inspection. Third, technically we could have used further text mining methods to improve the results. For this purpose we suggest to include other machine learning approaches in the future; in particular word2vec vector and correlated topic modeling (CTM). Fourth, we aim to perform a more traditional analysis with traditional metrics, including patents and license agreements, to verify the actual additional component of our approach and compare the results.

Finally, we need to implement a metric that aims to measure the actual impact of the knowledge presented by the company. Currently we only aim at the binary measure whether knowledge is transferred or not. It would be relevant to assess how important this specific knowledge is for the firm. This could enable a clear measurement of knowledge transfer contribution. (Nomaler & Verspagen 2008)

References

Agrawal, A. & Henderson, R. (2002), 'Putting Patents in Context: Exploring Knowledge Transfer from MIT', *Mgmt. Sci.* **48**(1), 44–60.

Agrawal, A. K. (2001), 'University-to-industry knowledge transfer: literature review and unanswered questions', *International Journal of Management Reviews* **3**(4), 285–302.

Ankrah, S. & Al-tabbaa, O. (2015), 'ScienceDirect Universities — industry collaboration : A systematic review', *Scandinavian Journal of Management* **31**(3), 387–408.

Ankrah, S. N., Burgess, T. F., Grimshaw, P. & Shaw, N. E. (2013), 'Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit', *Technovation* **33**(2-3), 50–65.

Argote, L. & Ingram, P. (2000), 'Knowledge transfer: A basis for competitive advantage in firms', *Organizational behavior and human decision processes* **82**(1), 150–169.

Arundel, A. & Bordoy, C. (2008), 'Developing internationally comparable indicators for the commercialization of publicly-funded research'.

Bekkers, R. & Bodas Freitas, I. M. (2008), 'Analysing knowledge transfer channels between universities and industry: To what degree do sectors also matter?', *Research Policy* **37**(10), 1837–1853.

Berry, M. W. & Castellanos, M. (2007), 'Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition', p. 241.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Branstetter, L. (2006), 'Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States', *Journal of International Economics* **68**(2), 325–344.

Brostroem, A. (2012), 'Firms' rationales for interaction with research universities and the principles for public co-funding', *Journal of Technology Transfer* **37**(3), 313–329.

Chapman & Hall/CRC (2010), *Handbook of Natural Language Processing, Second Edition.*

Cheah, S. (2016), 'Framework for measuring research and innovation impact', *Innovation* **18**(2), 212–232.

Cohen, W. M., Nelson, R. R. & Walsh, J. P. (2002), 'Links and impacts: the influence of public research on industrial r&d', *Management science* **48**(1), 1–23.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011), 'Natural Language Processing (almost) from Scratch', *The Journal of Machine Learning* **12**, 2493–2537.

D'Este, P. & Patel, P. (2007), 'University-industry linkages in the UK: What are the factors underlying the variety of interactions with industry?', *Research Policy* **36**(9), 1295–1313.

Drucker, J. & Goldstein, H. (2007), 'Assessing the Regional Economic Development Impacts of Universities: A Review of Current Approaches', *International Regional Science Review* **30**(1), 20–46.

Fernández-esquinas, M., Pinto, H., Pérez, M. & Santos, T. (2015), 'Technological Forecasting & Social Change Tracing the fl ows of knowledge transfer : Latent dimensions and determinants of university – industry interactions in peripheral innovation systems', *Technological Forecasting & Social Change* **113**, 266–279.

Franceschini, S., Faria, L. G. D. & Jurowetzki, R. (2016), 'Unveiling scientific communities about sustainability and innovation. A bibliometric journey around sustainable terms', *Journal of Cleaner Production* **127**, 72–83.

Franco, M. & Haase, H. (2015), 'University-industry cooperation: Researchers' motivations and interaction channels', *Journal of Engineering and Technology Management - JET-M* **36**, 41–51.

Geuna, A. & Rossi, F. (2011), 'Changes to university IPR regulations in Europe and the impact on academic patenting', *Research Policy* **40**(8), 1068–1076.

Gil Corrales, J., Song, W. & MacDonald, E. (2015), *Reproduction of Realistic Background Noise for Testing Telecommunications Devices*, Audio Engineering Society.

Griffiths, T. L. & Steyvers, M. (2004), 'Finding scientific topics.', *Proceedings of the National Academy of Sciences of the United States of America* **101 Suppl**, 5228–35.

Grimpe, C. & Hussinger, K. (2013), 'Formal and Informal Knowledge and Technology Transfer from Academia to Industry: Complementarity Effects and Innovation Performance', *Industry and Innovation* **20**(January 2016), 683–700.

Grün, B. & Hornik, K. (2011), 'topicmodels : An R Package for Fitting Topic Models', *Journal of Statistical Software* **40**(13), 1–30.

Gulbrandsen, M. & Slipersaeter, S. (2007), The third mission and the entrepreneurial university model, *in* 'Universities and Strategic Knowledge Creation: Specialization and Performance in Europe', chapter 4, pp. 112–143.

Heinze, N. & Hu, Q. (2006), 'The evolution of corporate web presence: A longitudinal study of large American companies', *International Journal of Information Management* **26**(4), 313–325.

Howells, J. (2002), 'Tacit Knowledge, Innovation and Economic Geography', *Urban Studies* **39**(5-6), 871–884.

Huggins, R., Johnstona, A. & Steffensonb, R. (2008), 'Universities, knowledge networks and regional policy', *Cambridge Journal of Regions, Economy and Society* **1**(2), 321–340.

Jensen, R. a., Jensen, R. a., Thursby, J. G., Thursby, J. G., Thursby, M. C. & Thursby, M. C. (2003), 'The Disclosure and Licensing of University Inventions'.

Jong, S., Barker, K., Cox, D., Sveinsdottir, T. & Van den Besselaar, P. (2014), 'Understanding societal impact through productive interactions: ICT research as a case', *Research Evaluation* **23**(2), 1–14.

Lin, J.-y. (2016), 'Technological Forecasting and Social Change Balancing industry collaboration and academic innovation : The contingent role of collaboration-speci fi c attributes', *Technological Forecasting and Social Change* .

Link, A. N., Siegel, D. S. & Bozeman, B. (2007), 'An empirical analysis of the propensity of academics to engage in informal university technology transfer', *Industrial and Corporate Change* **16**(4), 641–655.

Liyanage, C., Ballal, T., Elhag, T. & Li, Q. (2009), 'Knowledge communication and translation - a knowledge transfer model', *Journal of Knowledge Management* **13**(3), 118–131.

Lundberg, J., Tomson, G., Lundkvist, I., Skår, J. & Brommels, M. (2006), 'Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding', *Scientometrics* **69**(3), 575–589.

Malerba, F. (2007), 'Innovation and the evolution of industries', *Innovation, Industrial Dynamics and Structural Transformation: Schumpeterian Legacies* **23**(June 2004), 7–27.

Mao, W. & Chu, W. W. (2007), 'The phrase-based vector space model for automatic retrieval of free-text medical documents', *Data and Knowledge Engineering* **61**(1), 76–92.

Munari, F., Rasmussen, E., Toschi, L. & Villani, E. (2016), 'Determinants of the university technology transfer policy-mix: a cross-national analysis of gap-funding instruments', *Journal of Technology Transfer* **41**(6), 1377–1405.

Nomaler, Ö. & Verspagen, B. (2008), 'Knowledge Flows , Patent Citations and the Impact of Science on Technology Knowledge Flows , Patent Citations and the Impact of Science on Technology', **5314**(December).

Paukkeri, M.-s. & Honkela, T. (2010), 'Likey : Unsupervised Language-independent Keyphrase Extraction', (July), 162–165.

Polanyi, K. (1962), 'Personal knowledge: Towards a post-critical philosophy", chigago university press, chigago'.

Ponweiser, M. (2012), Latent Dirichlet Allocation in R, PhD thesis.

Rothaermel, F. T., Agung, S. D. & Jiang, L. (2007), 'University entrepreneurship: A taxonomy of the literature', *Industrial and Corporate Change* **16**(4), 691–791.

Salter, A. J. & Martin, B. R. (2001), 'The economic benefits of publicly funded basic research: a critical review', *Res. Policy* **30**(3), 509–532.

Schartinger, D., Rammer, C. & Fröhlich, J. (2002), 'Knowledge interactions between universities and industry in Austria: Sectoral patterns and determinants', *Innovation, Networks, and Knowledge Spillovers: Selected Essays* **31**, 135–166.

Siegel, D. S., Waldman, D. A., Atwater, L. E. & Link, A. N. (2004), 'Toward a model of the effective transfer of scientific knowledge from academicians

to practitioners: Qualitative evidence from the commercialization of university technologies', *Journal of Engineering and Technology Management - JET-M* **21**(1-2), 115–142.

Stahl, M. J., Leap, T. L. & Wei, Z. Z. (1988), 'Publication in Leading Management Journals As a Measure of Institutional Research Productivity', *Academy of Management Journal* **31**(3), 707–720.

Sung, T. K. & Gibson, D. V. (2000), 'Knowledge and Technology Transfer : Levels and Key Factors', *Proceeding of the 4th International Conference on Technology Policy and Innovation* .

Thursby, J. G. & Thursby, M. C. (2002), 'Who Is Selling the Ivory Tower? Sources of Growth in University Licensing', *Management Science* **48**(1), 90–104.

Tijssen, R. J. W., van Leeuwen, T. N. & van Wijk, E. (2009), 'Benchmarking university–industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities', *Research Evaluation* **18**(1), 13–24.

Toutkoushian, R. K., Porter, S. R., Danielson, C. & Hollis, P. R. (2003), 'Using publications counts to measure an institution's research productivity', *Research in Higher Education* **44**(2), 121–148.

Vincett, P. S. (2010), 'The economic impacts of academic spin-off companies, and their implications for public policy', *Research Policy* **39**(6), 736–747.

Xia, T. & Chai, Y. (2011), 'An improvement to TF-IDF: Term distribution based term weight algorithm', *Journal of Software* **6**(3), 413–420.

Zawdie, G. (2010), 'Knowledge exchange and the third mission of universities', *Industry & Higher Education* **24**(3), 151–155.

Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D. & Lu, J. (2016), 'Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research', *Technological Forecasting and Social Change* **105**, 179–191.

C CONFERENCE: AOM ATLANTA 2017

Title: *"Tracing Knowledge Transfer from Universities to Industry: A Text Mining Approach"*
Authors: Woltmann, Sabrina and Alkærsig, Lars
Conference: 77th Annual meeting of the Academy of Management
Year:*2017*
Organization: Academy of Management

Tracing Knowledge Transfer from Universities to Industry: A Text Mining Approach

ABSTRACT

This paper identifies transferred knowledge between universities and the industry by proposing the use of a computational linguistic method. Current research on university-industry knowledge exchange relies often on formal databases and indicators such as patents, collaborative publications and license agreements, to assess the contribution to the socioeconomic surrounding of universities. We, on the other hand, use the texts from university abstracts to identify university knowledge and compare them with texts from firm webpages. We use these text data to identify common key words and thereby identify overlapping contents among the texts. As method we use a well-established word ranking method from the field of information retrieval term frequency–inverse document frequency (TFIDF) to identify commonalities between texts from university. In examining the outcomes of the TFIDF statistic we find that several websites contain very related and partly even traceable content from the university. The results show that university research is represented in the websites of industrial partners. We propose further improvements to enhance the results and potential areas for future implementation. This paper is the first step to enable the identification of common knowledge and knowledge transfer via text mining to increase its measurability.

Keywords:

Text mining, knowledge transfer, impact assessment, university-industry

INTRODUCTION

Universities, as publicly funded institutions, conducting and disseminating research, are highly valued contributors to the knowledge development for economic growth and development (Feller, 1990; Howells, Ramlogan, & Cheng, 2012). The dissemination of their research outcomes is next to teaching seen as a major part of the impacts universities provide for their environment. In particular, the contribution of university research to economic development by fostering innovation leading to increased competitive advantages for industries and firms is today widely accepted (Cohen, Nelson, & Walsh, 2002; Huggins & Johnston, 2009). Academics and policy makers have in the past decades shown increasing interest in the identification of impact of the dissemination of university research; driven by the desire to ensure optimal allocation of limited public funding (Drucker & Goldstein, 2007; Rothaermel, Agung, & Jiang, 2007). Justification for the utilization of public funds thus became an incentive and are increasing the pressure to provide evidence for the return on public investments, so their societal and economic benefits are increasingly evaluated (O'Shea, Chugh, & Allen, 2008).

The increase of understanding and the evaluation of university research impacts became a political incentive and particularly the aspects of knowledge creation and transfer are in focus of assessments and evaluations (Agrawal, 2001).

Due to the high relevance of the topic, we aim to deepen the understanding of the economic impacts of university research dissemination by contributing with a new indicator and an additional novel measurement. Considering the current framework, this study takes a step back and aims to revive the work on the foundation of university research impact assessments: the notion of knowledge transfer.

The main objectives of our study are (i) to develop an additional measure of knowledge transfer (ii) to evaluate this new method using a case study of the Technical University of Denmark (DTU), chosen due to high accessibility of data, and (iii) to compare two different approaches of creating a relevant sample representing firm knowledge.

We seek to contribute by using computational methods, which are based on data mining processes, to develop our understanding of whether university knowledge is transferred and applied outside of formal collaboration and communication. The main method is derived from the field of natural language processing (NLP) and based on a concurrent text mining technique (Paukkeri & Honkela, 2010).

Text mining enables a trace from university research output, in form of publications, to corporate websites, annual reports or similar texts that give insight into firms' innovations, products and services. The goal is to identify correlations between these two types of texts, using this as an indicator for the transfer of new knowledge from the university to the firm. This paper should be seen as a first step towards identifying and understanding the characteristics of common knowledge between university and the industry. Our study contributes to the research on university-industry knowledge transfer by identifying correlations between university knowledge and firms commercially displayed knowledge via text analysis. We aim to increase insights into areas of common knowledge and mutual interests between universities and companies.

ECONOMIC IMPACT AND KNOWLEDGE TRANSFER

An extensive body of literature is concerned with the economic impact of university research. Since not one domain embraces all relevant aspects of this field of study, it has developed into a body of highly interdisciplinary works (Teixeira, 2014), providing a multitude of perspectives and definitions within the literature. Given the research diversity on publicly funded research impacts, today's understanding is comparatively well developed (Cheah, 2016). However, due to the diversity of scholars within the field, the understanding of 'economic impact' is used in varying contexts encompassing different notions, perspectives and dimensions (Cheah, 2016). Overall findings indicate different levels of economic impact for firms, sectors or regions. The benefit of university-generated knowledge is not spread uniformly across firms and sectors and national contexts (Bodas Freitas, Marques, & Silva, 2013), but examination of literature reviews and most influential empirical works reveals that the significant economic benefits of public-funded research are widely accepted (Agrawal, 2001).

Many studies follow the concept that knowledge transfer from universities to the industry is one of the key aspects of universities impact on the economy (Agrawal, 2001; Perkmann et al., 2013). "*Evidence suggests that even knowledge transferred through the formal university technology transfer channel […], is quite significant."* (Agrawal, 2001, p. 285). The body of academic literature consists of various sorts of impact studies ranging from single case studies, focusing on individual universities, to regional or even national surveys (Drucker & Goldstein, 2007; Huggins & Johnston, 2009; Rosenberg & Nelson, 1993). These diverse studies provide a great variety of methodological approaches aiming to identify university research impact, including qualitative and quantitative approaches.

Qualitative works are often concerned with in-depth understanding of motivation for university-industry collaborations or forms and channels of knowledge exchange, or focus on single universities as case studies (Ankrah, Burgess, Grimshaw, & Shaw, 2013; Perkmann & Walsh, 2009; Rothaermel et al., 2007; Siegel, Waldman, Atwater, & Link, 2004).

Quantitative studies on the other hand often provide particular insights about knowledge generation and knowledge transfer from universities to companies (D'Este & Patel, 2007; Schartinger, Rammer, & Fröhlich, 2002). Indicators used in quantitative studies comprise, among others, number of (co)-publications, number of successful university spin-offs, university income through license agreements, research collaborations and patents (Agrawal, 2001; Crespi, D'Este, Fontana, & Geuna, 2011).

Particularly patents and license agreements are often data of choice for estimating the true economic value of scientific and technical research outcomes (Bodas Freitas et al., 2013; Thursby, Jensen, & Thursby, 2001). Patents and/or licensing agreements are employed to assess the magnitude of knowledge utilized by firms. However, patents, licensing agreements, co-publications and the like do not capture all forms of knowledge exchange by far. They are mainly the most used proxy indicators due to their availability and international comparability (Thursby & Thursby, 2002). However, these indicators face long-standing criticism as they fail to represent a coherent picture of relevant knowledge spillovers (Cohen et al., 2002; Schartinger et al., 2002) and might not represent all specific aspects of successful commercialization as already stated by Agrawal and Henderson (2002). These indicators alone fail to provide a truly comprehensive picture of the knowledge contribution to the economy and yet the literature is dominated by those traditional measurements. Finding more holistic approaches for quantitative impact assessments of knowledge transfer from universities remains a great challenge.

Given these limitations we aim to provide a first step towards a novel measure that is applicable on a single case basis, which provides in-depth understanding like many qualitative studies do and is at the same time an additional quantitative approach, which provides generalizable and comparable results. We propose a computational linguistic approach for this purpose. The goal is to improve the detection of knowledge transfer without focusing on commercialization's, patents or the formal channels of knowledge transfers. The objective is to verify additional data sources and provide potential new indicators for tracing knowledge transfers from universities to the industry or vice versa.

METHODOLOGY

To compare our text samples from the university (DTU) and its partner or related firms, we chose well-established text-mining methods. Using these methods, we aim to identify new patterns of knowledge transfer, which are undetectable by existing indicators. The general assumption is that not all knowledge is necessarily patented or licensed, but it might be displayed in other texts formats. Hence, we use a method that statistically aims to detect word patterns in texts to identify textual pairs that represent the same or similar knowledge.

The applied method is based on the so-called 'bag of word assumption', which presumes that the words' order in a given document is irrelevant for the statistical analysis. Thus, the order of words in a given document is not taken into consideration and is treated as a set of independent features. Obviously, a document with unordered words will surely not express the same message as an ordered one and the features are by no means totally independent, as particular terms tend to occur more often in the particular documents. Furthermore, these methods assume that documents within a corpus are interchangeable and ordering of the documents in a corpus can be disregarded (Blei, Ng, & Jordan, 2003; Hofmann, 2001). However,

these assumptions do not entail any presupposition about for instance the independence or an identical distribution of the variables. The models operate in the space of distributions over words. Typically, documents are represented as feature-vectors, where a feature corresponds to one word (1-gram) or an ordered combination of words (bi- grams, ..., n-grams) (Berry & Castellanos, 2007). In this study, we focus solely on 1-grams, which limits the analysis because bi-grams like 'home made' or 'top ten' are divided in their single components and not identified as contextual unit.

Document-term matrix

The most common vector space representation of a document corpus is a document-term matrix, which contains feature (terms) frequencies associated to each document. Their rows correspond to documents and their columns to terms. The motivation is to achieve a representation of frequencies of semantically and contextual significant terms (Merritt, 2010). These matrices are commonly highly dimensional and sparse matrices (Berry & Castellanos, 2007). There are various schemes for determining the value that each entry in the matrix can take, depending much on the models used (Salton 1988).

In a term-document matrix, the element at (i,j) is the word count (frequency) of the i'th word (t) in the j'th document (d):

$$Document - Term\ Matrix = \begin{matrix} \\ d_j \end{matrix} \begin{matrix} t_i \\ \begin{pmatrix} x_{1,1} & \cdots & x_{1,i} \\ \vdots & \ddots & \vdots \\ x_{j,1} & \cdots & x_{j,i} \end{pmatrix} \end{matrix}$$

Word count (frequency) is sometimes modified and weighted for a better representation of the relevant feature of each document. Common weighting schemes include:

- Binary weighting, representing whether or not a term occurs in a document;
- Term-frequency weighting (TF), based on the number of occurrences in a document;
- Term-frequency inverse document frequency weight (TFIDF), using TF but assigning higher weight to terms that occur only in a small number of documents.

In our case, we converted all the single text corpora into document-term matrices applying (normalized) TFIDF weighting.

We additionally applied additive filtering of words not relevant to the context of a document by completely removing words that would occur in more than a certain percentage of documents in a corpus. The percentage was arbitrarily adjusted according to the method used, by assessing the outcome of the models and adjusting until obtaining satisfactory results.

TFIDF

This method is a numerical method used in various contexts and applied in text mining to calculate an order of content relevant words for documents. It is applied for text classification, summarization or content identification (Zhang et al., 2016). In order to identify commonalities between two documents, we used the TFIDF indexing to determine most characteristic words per document. These words can be regarded as key words describing the content of a document. The TFIDF indexing increases the value of the most relevant features of each document and devalues the feature occurring in more than a few documents.

TFIDF does not account for any synonymy or similarity and is purely bound to individual words, identifying only limited concepts of texts.

Different weighting calculations are possible for TFIDF indexing, but we opted for the most common weighting scheme, which additionally provides some normalization due to the included log transformation. For $t_i \in d_j$,

$$tf(t_i, d_j) = \sum t_i$$

We further have

$$idf(w, D) = \ln\left(\frac{N}{|\{d \in D: w \in d\}|}\right)$$

With N: Total number of documents and $|\{d \in D: w \in d\}|$: number of documents containing the word w. Finally, the TFIDF is obtained with the following multiplication:

$$tfidf(w, d, D) = tf(wd) \times idf(w, D)$$

We found that the representation of the keywords per document was improved for our comparison purposes, when performing the calculation on two separate corpora coming from two different sources. Both text sources do not have the same writing style. On one hand, websites contain a lot of spoken language and noise around the actual information. On the other hand, abstracts from publication papers are dense literature language. Hence, we chose this unusual approach of having two separate corpora for key word extraction.

Obviously, certain similarity measures could not be applied due to the two instances of word score calculation. We decided to include a maximum of 50 highest scoring terms per document. Reducing the dimensionality of documents to a binary list of maximal 50 terms enabled a comparison of keyword lists with each other. The TFIDF is a comparatively basic method, but is computationally economical and gives proficient results for any further analysis. Especially with short abstracts texts, the TFIDF keyword retrievals often resulted in lists shorter than five words, which needed to be considered for the later comparison.

Jaccard Similarity Coefficient

For the similarity measure between the two sets of identified keywords found thanks to the TFIDF, we used the Jaccard similarity coefficient as the metric. It is a statistic used for

measuring sets similarity. The Jaccard similarity is the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, one indicating most similarity (identical sets) and zero indicating least similar (no common feature in the two sets).

Given the set of keywords from one document of the publication database denoted K_A and the second set of keywords from one page of the websites denoted K_B, the Jaccard similarity denoted $J(K_A,K_B)$ is obtained with:

$$J(K_A, K_B) = \frac{K_A \cap K_B}{K_A \cup K_B} = \frac{K_A \cap K_B}{|K_A| + |K_B| - |K_A \cap K_B|}$$

We chose this similarity measure as it only includes occurrence and leaves order or values aside. The advantage is the low computational expense. This makes it attractive for a basic similarity assessment, which can of course be refined, by applying additional similarity measures to find more accurate matches.

The thresholds for a minimum similarity chosen for further examination were chosen based on brief manual investigation; meaning that we would only consider keyword lists with minimum Jaccard similarity values relevant enough for the manual inspection and potential matching. However, we observed that the Jaccard similarity tends to give better scores to small sets. For example, a 2 words intersection out of two sets of 3 words gives a very high Jaccard similarity (0.5) but is probably not indicating more related content than a 25 words intersection out of 50-words sets (0.33). Hence, we decided to set a common threshold to a minimum of 0.13 and another used indicator threshold consisted in multiplying the Jaccard index with the intersection of the two sets, giving higher weight to sets with a large intersection (higher amount of common words). The number of common words was multiplied with their Jaccard Similarity and needed to exceed 0.15×7, representing approximately 7 words intersection with Jaccard index of 0.15, approximately 7 common words out of 26-words sets. Thus, set pairs with Jaccard

Similarity lower than 0.15 need an higher than 7-word intersection in order to pass the criteria, while set pairs with Jaccard Index higher than 0.15 can have a lower than 7-word intersection in order to pass the matching criteria.

SAMPLES

The next section outlines steps undertaken for the generation of the text samples. The outline is divided into the generation of the text collections, representing university and industry knowledge and to identify common knowledge.

This study is using the case of the Technical University of Denmark (DTU) as scope of the study. Two main data sources are used in this study.

The first source is the university publication database named Orbit. The data set, provided by Orbit, contains a collection of research publication abstracts. These abstracts present main research outputs by employees of the DTU between 2005 until 2016. The database provides, among other information titles, keywords, author information and in most cases abstracts. Given the challenges to obtain a comprehensive sample of full text publications, abstracts were chosen as proxy of the universities research output, although this will not reflect the complete output.

The second data source, giving information on company knowledge and innovations, was gathered from firm websites. Selection criteria for the companies were (i) an English version of at least part of the website, (ii) a national branch of the company, and (iii) at least one common partner with the university.

Following these criteria the sample was produced using a hyperlink network from the university to its partners and partners of partners.

Publication Database (Orbit)

The selected data set from Orbit included all entries from January 2005 until August 2016, which resulted in a total of 76,627 publication entries. Of these entries, 43,745 included a full abstract, which were then categorized by research area and combined accordingly into separate corpora. This division of fields improves the later statistical analysis by dividing meaningful subsets for the data structure. Furthermore, computation time is reduced if a measure is only applied to smaller subsets of the data. The division resulted in 24 separate fields, which were aligned to department codes, provided by the database. Three of these subfields were irrelevant for the academic output of the university: (i) Publications registered to the university administration, (ii) publications registered to the bachelor program, and (iii) one set that was directly linked to a large company (this might have biased the findings significantly as the firm is directly involved in several hundreds of specially dedicated publications).

The remaining 21 fields are Electrical Engineering, Management Engineering, Physics, Compute, Chemistry, Mechanical Engineering, Environmental Engineering, Energy Conversion and Storage (EngConSto), National Food Institute, Nuclear Technologies, Aquatic Resources, Photonics, National Space Institute, Micro and Nanotechnology, Biochemistry, National Veterinary Institute, Civil Engineering, Wind Energy, Transport, Biosystems and Diverse[1].

These corpora will in the following be referred to as 'academic' corpora or by their individual name in case this is relevant for the interpretation of the results.

Firm Webpages

To identify the relevant firms for the firm based sample, we generated a simple directed network based on the relationships of the university with companies. A first network was

[1] This corpus contains publications, which do not fall under any of the above-mentioned categories.

generated on the basis of hyperlinks between webpages using the university as point of departure, (denoted Sample A). While an additional network was generated using university contracts to identify collaboration partners of the university and their partners (denoted Sample B). All companies connected to the university via hyperlinks and their direct partners were identified and stored, which resulted in a directed un-weighted second-degree network. The identified pages were downloaded and stored as HTML files.

The collected files were subsequently scanned for a Danish firm registration number and added to the text samples only if one was found for each given website. In a following step, the language of the page or the subpages was verified and only the English[2] content was stored. The online text samples were collected during August 2016 and September 2016[3]. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample, to avoid unnecessary pages and unrelated hyperlinks. In Denmark, universities are registered as companies and therefore have a Company registration number (CVR); so they had to be manually excluded.

Sample A

The first network contained 177 nodes, which represent individual company websites. These are connected to the university within a range of a path length of two, meaning that each node is either directly or over a common partner connected to the university page. The hyperlink network shows clear tendency to build clusters and it has some particularly central nodes. The nodes, which are highly interconnected and central for the structure of the network are mainly

[2] Danish firms provide a great amount of their information in English and the academic abstracts are in English, which enables a comparison, based on keywords between Danish firms and Danish university research in English.

[3] The script used to identify and download the pages can be found at https://github.com/nobriot/web_explorer

online service platforms, including transportation and types of yellow pages and firm registries. The texts from this network contain overall around 120,000 unique terms. We assigned each text to its website URL, resulting in 121 single text corpora based on individual websites (with up to 1000 webpages). 56 smaller websites had less than 5 pages after language filtering and were combined to one single corpus, as these would be too small to apply the relevant statistical analysis, as they are mainly composed of brief introduction pages of the home pages, not containing any relevant information.

During the network generation it became apparent that many official partners are not necessarily connected with a hyperlink to the university main pages. We included the web Sample B to account for this.

Sample B

To generate an additional sample another network was created based on Danish companies with a formal connection to the university, namely a collaboration contract. Hence, we commenced building the second network with around 686 first-degree firms, which had a contract with the university between the years 2013 and beginning of 2016. Those new websites were collected and their online partners were also identified. This generated a fully new network including more content related companies. The identified firms operate mainly in technology intensive sectors and are firms with strong R&D divisions.

The second network contained 686 nodes and of which 312 were identified as Danish companies. This sample, resulted in 243 single text corpora, based on individual websites (with up to 1000 webpages) and an additional corpus again containing 69 smaller pages. For the later analysis we will refer to the sample that is solely based on hyperlinks as sample A and the sample including internal contract information as sample B.

Pre-processing

Text pre-processing describes the task of converting unstructured raw text into an order of computationally and statistical useful and linguistically meaningful units. The pre-processing is an essential part of any text analytical procedure, since the characters and words are identified at this stage as the units passed on to further text mining stages (Paukkeri & Honkela, 2010).

Pre-processing of text, which is also known as tokenization includes in our case the following steps:

- Define word boundaries as white spaces.
- Remove unessential elements (e.g. coding tags, punctuation, and numbers).
- Convert all characters to lower case (makes the identification of abbreviations challenging).
- Strip the texts from additional white spaces.
- Remove stopwords, meaning most frequent words, which do not carry content information (in some cases, topic specific stopwords were added).
- Apply stemming which is beneficial to merge the inflected word forms into the corresponding stem.

Results of this pre-processing revealed some challenges especially for the academic abstracts. For instance chemical formulas and similar notations rely on numbers, short abbreviations and punctuation. So after pre-processing the only possibility to identify the concurrent formulas would be the prospect that the removal of numbers and punctuation results in the same string in both types of texts that can be seen as an equivalent to a term representation of the formula. Additionally, some very specific abbreviations are sometimes hard to identify, meaning that the results of the tokenization does not seem to make much sense, but are actually

describing very particular features of some publications (e.g. omniitox, which is name of a European project, or 'modelpbpk', standing for PBPK modeling). Finally, we merged the pre-processed texts into text corpora, which are a large ordered set of documents, to ensure structured sets of texts.

RESULTS

The described TFIDF indexing was used to assess the documents' similarity. We divided the results into the two web data samples for illustration. The results vary greatly due to the high diversity of the text corpora from the firm samples. After all pre-processing steps the sample A encompassed 117 websites containing 30,241 single pages and sample B with 243 websites and 77,421 pages.

We classified the found text pairs or matches into 5 main categories:

- 1st order: Web texts which are related to a university publication
- 2nd order: Web texts which are very likely to be related but miss an actual clear link
- 3rd order: Web texts which clearly come from the same area, but concern a different sub-field of the area
- 4th order: text pairs that contain similar topics but there is no deeper connection
- 5th order: text pairs with no overlap at all.

It has to be remarked that the pairing of the web text files and the abstracts resulted in several recurrent hits, meaning that the overall number of different pairs is significantly lower than the raw found matches, due to the fact that companies often display the same text content on more than one page. However, still one page could have several hits, so we excluded pairs, which represented the same website and the same abstract, but a different page from the website.

We decided to perform the manual investigation on the original texts, without any pre-processing to ensure that the actual content of the documents was understood.

TFIDF results for the academic corpora

The application of the TFIDF indexing on the 21 academic corpora resulted in a given set of key words for each document. Several academic expressions were hereby filtered out and context relevant words were identified. Table 1 shows the 5 most relevant words for each university department.

Insert Table 1 about here

These words represent the content of the departments satisfactorily considering the exclusion of too recurrent words. A manual inspection of the sample confirmed an adequate representation of keywords on corpus (departments) and document level. However, the collected abstracts were relatively short (4-6 sentences), which limited the content and representation of keywords per se. The same comprehensiveness of presentation of keywords accounts for the websites.

Results of the Comparison with Sample A

In the following we compared each keyword set from any website with the keywords of each abstract in every academic corpus. This led overall to 1,306,139,031 comparisons. For the chosen threshold for the Jaccard similarity (see Methodology section), 385 document pairs were considered as matching documents (including all pairs). The matching rate of relevant pairs was 2.9×10^{-5} %. The highest scoring pair reached 0.235 Jaccard Similarity representing in our case 19 common words out of 81 total keywords. As a benchmark, calculating the Jaccard Similarity

between the different abstracts within the academic corpora, given the same threshold, the threshold was exceeded more than 0.009% of the time. The highest Jaccard similarity was in this case close to 1. Showing clearly that the academic corpora documents are found to have more in common among each other than with the sample A.

The average Jaccard similarity for matches in the sample A was 0.125, which is rather low. Only 22 pairs exceeded 0.15 Jaccard similarity. The identified pairs were in the following manually examined. . Highest Jaccard similarity scores were dominated by a word co-occurrence of country names, which is likely to be only of limited contextual relevance. Additionally, some text pairs were identified as similar due to a common foreign language, which was detected in both texts like for instance parts of German or Danish. Indeed many similar pairs, show that the dominating attributes were country names, but that among the top ten pairs were some in which the common words with more content relevance as shown in Table 2.

Insert Table 2 about here

With a manual inspection of the found pairs we found a limited number of common contents and DTU related research content. We found the following classifications:

- 1st order matches: 4
- 2nd order matches: 2
- 3rd order matches: 10
- 4th order matches: 4
- 5th order matches: 5

There were no 1st order or 2nd order pairs identified below the Jaccard Similarity threshold of 0.130. It should be mentioned that this sample contained a considerable number of

1st order pairs (16), which were representing websites of public entities, which in some cases were even part of the university itself. Hence, these pairs were subtracted from the overall 1st order pairs. However, these were correctly identified pairs. The overall correct identification would therefore be 20 correct identified pairs. Eliminating the country pairs and hits under 0.130 we have 41 relevant pairs left and the 1st and 2nd order pairs were 53.65% from the overall findings. The common contents were mainly related to system inventions, or presentations given by DTU employees and mentioned on the respective websites.

Comparison of Sample B

For each page per website of sample B, we calculated the keywords via the TFIDF indexing and compared with the academic keyword sets. In the case of sample B this accounted for 3,343,890,411 compared pairs and 974 of them passed the chosen threshold. This is again a percentage of 2.9×10^{-5} % found pairs, which is identical to sample A's matching rate. This resulted in 25 text pairs scoring a Jaccard Similarity over 0.15 but none over 0.18, which is lower than Sample A's result. The average Jaccard similarity was 0.121 for found matches, which was lower than the one from sample A.

Most common words were more diverse than the ones of sample A. The resulting matches of keywords consisted of words that have more content relevance, however the highest pairs are still consisting country related words (refer to Table 3).

Insert Table 3 about here

The manual verification of the text pairs revealed that the matches scoring under 0.130 Jaccard similarity are definitely less relevant and contain mainly 4th order pairs than the pairs that

exceed this threshold. After removing all pairs under 0.130 Jaccard Similarity and excluding the country pairs we were left with 89 relevant pairs. We identified the following numbers for the classes of the text pairs:

- 1st order matches: 13
- 2nd order matches: 10
- 3rd order matches: 22
- 4th order matches: 23
- 5th order matches: 16

This means that 27.38 % of the matching pairs were clear references to the university knowledge or were highly likely related. We had 5 pairs (5.95%), which we could not clearly classify as the information provided by the abstract was too limited, or the content too specific and would require an expert opinion of the specific field. Only 19.05% were pairs that have no overlap and were wrongly identified.

DISCUSSION

Generally is evident that the results from sample A and B vary in their quality (text content) and quantity. The most relevant matches 1st and 2nd order describe clearly the use of common, partly by the university invented methods and their direct application. Three of the websites state the university as source of these methods or tools. Some of the matches are towards the same website but identify different contents, so one site is responsible for 4 of the 1st order matches. Within the 1st order we found one match where the company that does display the content refers to another company with which the university has the topic related contracts and the content matched extremely well. In other cases, parts of the actual abstract are directly quoted, but without a clear reference to the university.

The 2nd order pairs show often a strong overlap in scope content and used methods, but lack a clear verification or linkage to the university, which the 1st order pairs contain.

Sample A's 1st order pairs were mainly a clear display of research results either on the pages of other public entities, conference summaries or similar. Resulting in the identification of clear related, but in terms of commercial use and knowledge transfer maybe not very relevant. Sample B's 1st order pairs are dominated by the use of university developed tools and models and are therefore extremely relevant in terms of our research objectives.

Given that sample A is a sample containing mainly websites that are not related to any of the university's research this is a positive outcome, as it verifies that the method finds communalities where there are some present. Generally, the performance of this simple measure is comparatively successful as it succeeds in identifying knowledge overlaps.

A further confirmation is the significantly higher number of commonalities among the academic keywords than between websites and academic corpora, even though they refer often to different topics, especially since a technical university as such has a great overlap among the research fields. In sample A, many pairs were correctly identified but the identification of purely private enterprises was not impeccable. The comparatively small number of 1st and 2nd order pairs show that there would be additional identification mechanisms suitable to obtain more results. However, it shows that the pairing can identify the use of university related knowledge and even the use of university created knowledge.

The high number of 3rd and 4th degree order in sample B represents companies that use the common contents like particular models, instruments, or metrics in the same or closely related fields, but are rather unlikely connected to the university's research.

The performance of the TFIDF indexing, especially given the benchmark comparison matching between the different academic corpora, shows that it identified 186.414 pairs that reach the threshold even though the abstracts are significantly shorter than the webpages, which means the quantity of text for matching is reduced and findings should be less. Some more trails to find optimal thresholds need improvement and additional randomized testing is necessary, but the results are promising.

CONCLUSION

This study provides a first attempt to develop an additional measure of knowledge transfer by using texts as main data sources. Our test case shows that the identification of university knowledge in firms' websites is clearly possible by applying the given statistical measures. We examined two different samples of websites and our results suggest that our approach does work for formal as well as for informal or second-degree partners of the university. The overall outcome identifies common grounds between companies and the university.

We can identify texts that show on the one hand either a clear relation to university knowledge and furthermore identify the companies that deal with very related topics. This can be used to identify the universities knowledge transfer and additionally most common areas of interests from universities and companies. We see this as a great step towards the actual detection of knowledge spillovers and transfer, even though it is certainly just an addition to current metrics.

Limitations

The text samples of firm websites for the study are not exhaustive as especially PDF formats and similar were not yet included in the sample. Additionally an additional identification

of Danish firms would be beneficial. Regarding the representation via abstracts of publications must be said that the availability of full text would have been beneficial especially since the content of academic abstracts is per se very limited.

Finally, the TFIDF indexing is a rather simple method, which is incapable to capture contexts, meaning that in case different words are used to describe the same subject this method would fail to identify a connection.

Future research

Next steps for the improvement of this approach are to increase the quality and quantity of the text data, by gaining access to full text publications and potentially annual reports from relevant firms. For future research we also aim to provide automated classifications into the 5 classes, which will only have to be verified by humans to decrease the amount of manual labor. We aim to combine our approach it with additional statistical approaches to increase the performance. Concurrent machine learning approaches will come in handy and enable us to enhance the current results. Ideally we will be able to test our next results against the outcome of traditional metric.

REFERENCES*

Agrawal, A., & Henderson, R. (2002). Putting Patents in Context: Exploring Knowledge Transfer from MIT. ***Management Science***, 48(1): 44–60.

Agrawal, A. K. (2001). University-to-industry knowledge transfer: literature review and unanswered questions. ***International Journal of Management Reviews***, 3(4): 285–302.

Ankrah, S. N., Burgess, T. F., Grimshaw, P., & Shaw, N. E. (2013). Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit. ***Technovation***, 33(2–3): 50–65.

Berry, M. W., & Castellanos, M. (2007). ***Survey of Text Mining II: Clustering, Classification, and Retrieval***. London: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. ***Journal of Machine Learning Research***, 3(4–5): 993–1022.

Bodas Freitas, I. M., Marques, R. A., & Silva, E. M. D. P. E. (2013). University-industry collaboration and innovation in emergent and mature industries in new industrialized countries. ***Research Policy***, 42(2): 443–453.

Cheah, S. (2016). Framework for measuring research and innovation impact. ***Innovation***, 18(2): 212–232.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and Impacts : The Influence of Public Research on Industrial R & D. ***Management science***, 48(1): 1–23.

Crespi, G., D'Este, P., Fontana, R., & Geuna, A. (2011). The impact of academic patenting on university research and its transfer. ***Research Policy***, 40(1): 55–68.

D'Este, P., & Patel, P. (2007). University-industry linkages in the UK: What are the factors underlying the variety of interactions with industry? ***Research Policy***, 36(9): 1295–1313.

Drucker, J., & Goldstein, H. (2007). Assessing the Regional Economic Development Impacts of Universities: A Review of Current Approaches. ***International Regional Science Review***, 30(1): 20–46.

Feller, I. (1990). Universities as engines of R& D-based economic growth: They think they can. ***Research Policy***, 19(4): 335–348.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. ***Machine Learning***, 42: 177–196.

Howells, J., Ramlogan, R., & Cheng, S. L. (2012). Innovation and university collaboration: Paradox and complexity within the knowledge economy. ***Cambridge Journal of Economics***, 36(3): 703–721.

Huggins, R., & Johnston, A. (2009). The economic and innovation contribution of universities: A regional perspective. ***Environment and Planning C: Government and Policy***, 27(6): 1088–1106.

Merritt, D. (2010). ***Adventure in Prolog***. New York: Springer

O'Shea, R. P., Chugh, H., & Allen, T. J. (2008). Determinants and consequences of university spinoff activity: A conceptual framework. ***Journal of Technology Transfer***, 33(6): 653–666.

Paukkeri, M., & Honkela, T. (2010). Likey : Unsupervised Language-independent Keyphrase Extraction. ***Proceedings of the 5th International Workshop on Semantic Evaluation***: 162–165.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., et al. (2013). Academic engagement and commercialisation: A review of the literature on university-industry relations. ***Research Policy***, 42(2): 423–442.

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: Impacts of university-industry relations on public research. ***Industrial and Corporate Change***, 18(6): 1033–1065.

Rosenberg, N., & Nelson, R. R. (1993). American universities and technical advance in industry. ***Research Policy***, 23: 323–348.

Rothaermel, F. T., Agung, S. D., & Jiang, L. (2007). University entrepreneurship: A taxonomy of the literature. ***Industrial and Corporate Change***, 16(4): 691–791.

Schartinger, D., Rammer, C., & Fröhlich, J. (2002). Knowledge interactions between universities and industry in Austria: Sectoral patterns and determinants. ***Innovation, Networks, and Knowledge Spillovers: Selected Essays***, 31: 135–166.

Siegel, D. S., Waldman, D. A., Atwater, L. E., & Link, A. N. (2004). Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: Qualitative evidence from the commercialization of university technologies. ***Journal of Engineering and Technology Management***, 21(1–2): 115–142.

Teixeira, A. A. C. (2014). Evolution, roots and influence of the literature on national systems of innovation: A bibliometric account. ***Cambridge Journal of Economics***, 38(1): 181–214.

Thursby, J. G. J. J. G., Jensen, R. a., & Thursby, M. C. M. (2001). Objectives, characteristics and outcomes of university licensing: A survey of major US universities. ***The Journal of Technology Transfer***, 26(1): 59–72.

Thursby, J. G., & Thursby, M. C. (2002). Who Is Selling the Ivory Tower? Sources of Growth in University Licensing. ***Management Science***, 48(1): 90–104.

Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. ***Technological Forecasting and Social Change***, 105: 179–191.

TABLE 1*

Most relevant words for each DTU department

Department (corpus)	Most relevant words				
Compute/Math	attack	secur	graph	network	code
ChemBiochem	enzym	polym	membran	oil	catalyst
Chemestry	hydrogen	zeolit	liquid	membran	hydrogen
CivilEng	solar	crack	collector	moistur	stress
ElectEng	antenna	convert	fault	robot	flow
EngConSto	magnet	membran	carbon	anod	field
EnviEng	landfil	sludg	methan	bioga	climat
MAN	servic	network	materi	configur	risk
MechEng	weld	stress	steel	wind	bear
MicroNano	magnet	graphen	cantilev	laser	reson
PhotoEng	quantum	thz	dispers	data	convers
Physics	nanoparticl	pbri	water	mode	plasma
BioSys	biofilm	peptid	resist	dna	aeruginosa
Transport	brbr	til	der	ship	capac
Wind	ref	composit	instal	fibr	accord
Food	efsaq	claim	substanc	salmonella	vitamin
Aqua	egg	prey	migrat	codend	genet
Space	burst	graviti	mcrab	cluster	nustar
Nuc	msupsup	neutron	iodin	supsupi	risø
Vet	resist	serotyp	intestin	fmdv	genotyp
Diverse	magnet	film	grain	turbin	electrod

TABLE 2*

Word co-occurrence for Sample A

Words and their co-occurrences (100 top-words)							
latvia	103	research	23	dtu	10	phospholipid	8
hungari	103	slovakia	23	factor	10	enzym	8
cyprus	102	support	18	phospholipas	10	die	7
bulgaria	102	sweden	17	ischem	10	experiment	7
lithuania	100	renew	16	european	10	plant	7
estonia	99	electr	16	fibril	10	digest	7
finland	91	technolog	15	atrial	10	nation	7
greec	88	risk	14	obes	10	qsar	7
slovenia	88	energi	14	industri	10	procedur	7
czech	88	student	14	stratif	10	sustain	7
republ	88	grid	13	cost	9	ist	7
romania	81	ion	13	physic	9	microscopi	7
technic	49	consumpt	13	fuel	9	knowledg	7
univers	47	fast	13	young	8	databas	7
denmark	41	scatter	12	hydrolysi	8	dynam	7
engin	39	thomson	12	austria	8	und	7
electron	38	collect	12	den	8	countri	7
list	37	power	12	emiss	8	interest	7
sourc	36	der	11	earth	8	properti	7
issu	33	wind	11	liposom	8	ein	7
publish	33	suppli	11	comment	8	von	7
depart	32	gas	11	member	8	programm	7
note	32	learn	10	secretori	8	pretreat	7
luxembourg	26	coronari	10	netherland	8	specif	7
ireland	24	myocardi	10	bioga	8	storag	7

TABLE 3*

Word co-occurrence for Sample B

Words and their co-occurrences (100 top-words)											
electr	91	heat	39	properti	28	ist	24
cycl	84	caus	39	market	28	amplifi	24
fuel	80	sourc	37	spot	28	med	23
der	78	stress	37	damag	28	month	23
environment	70	den	37	document	28	des	23
impact	64	von	36	failur	28	sustain	23
solar	61	mit	35	coal	28	layer	23
die	61	advanc	34	das	28	countri	22
renew	60	werden	34	nois	28	som	22
life	60	depend	33	turbin	27	cost	22
assess	60	tension	33	review	27	consum	22
und	59	deform	32	econom	27	conduct	21
gas	55	analys	32	characteris	27	har	21
fossil	54	mass	32	fibr	27	produc	21
wind	53	emiss	32	obes	27	storag	21
lca	52	auf	31	gain	27	decis	21
temperatur	52	biomass	31	creat	26	electrochem	21
greenhous	49	calcul	30	figur	25	manag	21
power	49	degrad	30	til	25	equat	21
grid	48	energi	30	global	25	growth	21
für	47	mechan	29	resourc	25	sector	20
ein	47	consumpt	29	suppli	25	smart	20
demand	47	determin	29	technolog	25	index	20
plant	41	weld	29	denmark	25	ion	20
climat	40	futur	29	transport	25	averag	19

D CONFERENCE: SMS 2017

Title: *Company-University Collaboration Types As A Determinant For Knowledge Transfer*
Author: Woltmann, Sabrina and Alkærsig, Lars
Conference: SMS Special Conference Costa Rica
year: *2017*

Introduction

Becoming 'the most competitive and dynamic knowledge- based economy in the world' (Commission 2010)[p. 1] is part of the agenda of the European Union. Hereby declaring that investments into research and (technological) development (R & D) needs to be the focus to face the increasing global competition. Gross domestic expenditures in 2014 in R & D added up to 299 billion Euro(Eurostat 2017). Around 55% of these expenditures were made by business enterprises, while 32% were government funded.

Universities have traditionally key roles in national R & D infrastructures and are today key players for research driven inventions (Cohen et al. 2002). Government and/or EU financing incentives, like the EU Research Framework Programs[1] increased the interest and value of university-industry joint research projects. Companies use it to increase their expertise using external knowledge sources (Fabrizi et al. 2016, Azagra-Caro et al. 2009). Hence, interaction between firms and universities is a key aspect for research driven innovation todays industries.

Changes in the last decades in regulations and policies increased collaborative activities of universities and companies (Geuna & Rossi 2011). Most fostering aspects are the intensified public funding, available tax reduction schemes, changes in the intellectual property rights (IPR) and access to vast interdisciplinary research facilities at the universities (Lissoni et al. 2009, Geuna & Rossi 2011, Munari et al. 2016). Relationships and collaboration between companies and universities differ greatly in their structures, intensity and quality, leading to varying outcomes and benefits for the companies (D'Este & Patel 2007). Knowledge transfer (KT) includes the acquisition and utilization of novel technologies or innovations. Among scholars most discussed are benefits from university collaboration when they generate either inventions that are commercialized, including licenses, royalties and patents or create any other outcome that is IPR protected (Arundel & Bordoy 2008, Crespi et al. 2011, Rothaermel et al. 2007) .

Since only a small proportion of KT is actually directly commercialized (Agrawal & Henderson 2002) an expansion of the measurement is indispensable. Thus, we propose an additional perspective to identify and verify outcomes of differently structured relationships between companies and universities. As companies often invest highly into university collaboration and by far not all investments lead to potentially commercializable innovations(Jensen et al. 2003, Cohen et al. 2002) we aim to assess whether the level of interaction changes the level of KT for the companies.

To identify the levels of KT we use new computational metrics: *text mining*. We trace patterns from texts related to university research, university publications, and texts from companies, company homepages, to identify commonalities. Novel statistical methods allow us to identify common content and therefore detect KT. We aim to show that companies

[1](European Comission 2016)

in qualitatively different relationships to universities can very well harvest commercially relevant outcomes, which are displayed on the companies online presences.

Theoretical framework

In order to be fruitful the collaboration between universities and companies rely on three main factors: a) university's dissemination activities, b) absorptive capacity of the company, and c) the type of the relationship between company and university. We focus solely on the two latter, to enhance understanding about the potential outcomes from company perspectives.

First, the absorptive capacity of a company, as originally defined by (Cohen & Levinthal 1990) is depending on a firms ability 'to recognize value of new, external information, assimilate it and apply it'. This basic understanding shows that the absorption of university research by a company might not depend only on their interaction, but also on the company‘s features. Therefore we aim to compare different companies in relationships with universities. Expecting that companies, with a higher expenditures in R&D and/or high prior knowledge benefit even from comparatively loose relationships. In the literature on university-industry relations, five different levels of formalization can be distinguished: i) longstanding formalized research collaboration [2],ii) medium and short term formalized research collaboration [3], iii) direct formal relationship not based on research related activities, iv) having a common partner with a university resulting in an indirect relationship, v) no traceable research or other ties, but confirmed geographical proximity to the university.

The **control variables** we propose for an even comparison are: the firm size (defined as number of employees), type of industry (using NACE codes) to account for low and high tech industries and the company employees educational level from the national Danish statistics bureau (Statistics Denmark).

For the comparison of the outcome we suggest to measure observable transfer of knowledge, since traditionally outcomes of university-industry collaboration are often measured in terms of KT(or technology transfer [4]). KT is used to assess the implications of formalized, informal collaboration (Nomaler & Verspagen 2008, Freitas et al. 2013). Two main distinctions between for KT can be made: formal and informal (Grimpe & Hussinger 2013).

Formal knowledge transfer is clearly defined by the outcomes of interactions, which are the result of direct formal ties (contracts) between

[2]We set the time frame for more than 3 years and/or recurrent contracting.

[3]We set a maximum of 3 years without subsequent contracting

[4]The term Knowledge transfer will in the following be used interchangeably for Technology transfer

the partners. Formal KT comprises all 'transfer mechanisms that embody or directly result in a legal instrumentality such as, for example, a patent, license or royalty agreement (Link et al. 2007). Current quantitative research focuses mainly on formal KT measurements, looking at the outcomes including profits generated by patents, spin-outs, royalty and/or licenses agreements(Grimpe & Hussinger 2013). Hereby the actual KT is not fully captured (Agrawal & Henderson 2002), because many joint activities do not results in an IP protected innovation, but the knowledge might still be transferred.

Informal knowledge transfer "is facilitating the flow of technological knowledge through informal communication processes, such as technical assistance, consulting, and collaborative research" (Link et al. 2007) [p. 642). It comprises any form of KT that does not imply a IPR regulated outcome and might not be the result of a formalized relation ship. Informal KT is not easily measured and is mainly identified via in-depth case studies (Broström 2012), but the actual outcomes are no quantitative measures.

Dependent variable: Observable knowledge transfer

Due to the limitations of the KT concepts we seek to include a new metric that allows us to capture any type utilized knowledge coming from a university independent from the transfer channel or formalization of the relationship. For the purpose of this study we define observable KT as KT that can be identified via text mining algorithms, which allow comparing university research outcomes (publications) with company online presences (websites, social media sites and annual reports). Commonalities in the content indicate a commercial use of research. The level of observable KT is ranked within the four quartiles of the given identification level of KT, which is defined by intensity of the expressed overlap between the contents. A company ranked in the first quartile (identified KT lower than 25%) has low KT, companies ranked in the 2nd and 3rd quartiles (from 25% to 75%) will be seen as intermediate and every company ranked in the 4th quartile (above 75%) will be considered having a high KT. All remaining companies are considered as having no observable KT.

Types of collaboration

We aim to identify the characteristics of companies that would suggest most higher observable KT taking into account the type of relationship. First we aim to assess whether the observable KT is related to the type of commitment in the company-university relationship.

> ***H 1a:*** *Longstanding collaboration, including large research projects or recurring collaboration, (potentially with formal KT) result in* ***high*** *observable KT for the collaborating company.*

However, given the varying absorptive capacity of companies, we suggest to diversify the picture by focusing on more lose relationships. Further differentiation between contract based connections is necessary to identify further potential benefits.

> ***H 1b:*** *Short-term research collaboration with no subsequent contracting result in* ***high*** *observable KT for the collaborating companies.*

> ***H 1c:*** *Short-term non research related contracts result in* ***high*** *observable KT for the collaborating companies.*

Consequently this model has to be extended to companies with no direct collaboration with the university. The notion of knowledge spillovers in proximity of universities (Drucker & Goldstein 2007, Arundel & Geuna 2004) lead to the following assumptions:

> ***H 2a:*** *Collaboration with companies that are collaborating with universities results in* ***medium*** *or* ***high*** *observable KT.*

> ***H 2b:*** *Companies, which are located in the proximity of the university, receive* ***low*** *or* ***medium*** *observable KT.*

Sample

We collected relevant text data representing a) company profiles and b) university research knowledge. We use publication data from the Technical University of Denmark (DTU), representing the research output and a collection of online texts from company websites.

Between 2006-2017 DTU had a total of 78,627 publications and provides 43,745 academic abstracts and 23,402 full-text publications. Relevant publications have to be co-authored by at least one member of the university and need to have English text. We divided the texts into 21 separate research fields including mathematics, biochemistry, chemistry, civil engineering, electrical engineering, energy conversion, environmental engineering, management, mechanics, nanotechnology, photonics, physics, biology, transport, wind energy, nutrition science, aquatics, space research, veterinary, nuclear technologies and one with diverse entries.

Companies with any type of contract between 2006 and 2016 were considered to be relevant, as well as partners of these companies (second degree partners). We identified 1256 companies and 768 second degree partners [5].Relevant websites have to display a Danish registry number of the firm (CVR number) and partly English content. The content was stored as HTML files.

[5]To identify indirect partner firms, we created a network based on hyperlinks between websites.

Methods

We use statistical tools from the field of natural language processing (NLP) to identify text correlation and similarities (Indurkhya & Damerau 2010, Collobert et al. 2011).We apply common **text pre-processing** steps, which convert unstructured raw text into statistical useful units (Paukkeri & Honkela 2010).

For pattern recognition we use **Term-frequency, inverse document frequency** (TFIDF) a simple numerical indexing method, which has proven to give promising results. It allows identifying the most relevant words by extracting the words most unique to a given text(Zhang et al. 2016).

For content identification we use **latent dirichlet allocation** (LDA), a fully automated method based on statistical learning. It identifies latent (unobservable) content structures (Blei et al. 2003, Griffiths & Steyvers 2004) and translates them into topics. These topics enable classification of text content.

We use word2vec (w2V) to further identify communalities between the texts and ensure computational optimal outcomes. It describes methods that are used to reconstruct the contexts of words by taking texts and producing a vector space. Word vectors are assigned by contexts and so related words are located in close proximity to one another(Rong 2014). We use this to identity strongly related words and texts. The combination of our methods ensures minimal manual work.

Conclusion

Our approach allows to gather a more coherent picture about the benefits of university collaboration for companies. The assessments of company-university collaboration outcomes have been focused on the measures of formal knowledge transfers (Salter & Martin 2001, Teixeira & Silva 2013, Jensen et al. 2003). This study has the potential to add a new perspective to current metrics and will open insights into the variation in outcomes of different types of collaboration. The study aims to achieve a more coherent understanding about the benefits and innovative potential of university-company interactions. This can be used by the industry to re-asses their engagements and activities. The results are also likely to influence the common view on university- company collaboration, as it can provide an additional measure for acquired and used knowledge obtained from a university.

References

Agrawal, A. & Henderson, R. (2002), 'Putting patents in context: Exploring knowledge transfer from mit', *Management science* **48**(1), 44–60.

Arundel, A. & Bordoy, C. (2008), Developing internationally comparable indicators for the commercialization of publicly-funded research.

Arundel, A. & Geuna, A. (2004), 'Proximity and the use of public science by innovative european firms', *Economics of Innovation and New Technology* **13**(6), 559–580.

Azagra-Caro, J. M., Carat, G. & Pontikakis, D. (2009), 'University-industry cooperation in the Research Framework Programme', *JRC Scientific and Technical Reports* pp. 1–8.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Broström, A. (2012), 'Firms' rationales for interaction with research universities and the principles for public co-funding', *The Journal of Technology Transfer* **37**(3), 313–329.

Cohen, W. M. & Levinthal, D. A. (1990), 'Absorptive capacity: A new perspective on learning and innovation', *Administrative science quarterly* pp. 128–152.

Cohen, W. M., Nelson, R. R. & Walsh, J. P. (2002), 'Links and Impacts : The Influence of Public Research on Industrial R & D', *Management science* **48**(1), 1–23.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011), 'Natural language processing (almost) from scratch', *Journal of Machine Learning Research* **12**(Aug), 2493–2537.

Commission, E. (2010), Lisbon strategy evaluation document, Technical report, European Commission, Brussels.

Crespi, G., D'Este, P., Fontana, R. & Geuna, A. (2011), 'The impact of academic patenting on university research and its transfer', *Research Policy* **40**(1), 55–68.

Drucker, J. & Goldstein, H. (2007), 'Assessing the regional economic development impacts of universities: A review of current approaches', *International regional science review* **30**(1), 20–46.

D'Este, P. & Patel, P. (2007), 'University–industry linkages in the uk: What are the factors underlying the variety of interactions with industry?', *Research policy* **36**(9), 1295–1313.

European Comission (2016), 'Home page - FP7 - Research - Europa'.
URL: *http://ec.europa.eu/research/fp7/index_en.cfm https://ec.europa.eu/research/fp7/index_en.cfm*

Eurostat (2017), 'R & D expenditure'.
URL: *http://ec.europa.eu/eurostat/statistics-explained/index.php/R_&_D_expenditure*

Fabrizi, A., Guarini, G. & Meliciani, V. (2016), 'Public knowledge partnerships in European research projects and knowledge creation across R&D institutional sectors', *Technology Analysis & Strategic Management* **0**(0), 1–17.

Freitas, I. M. B., Marques, R. A. & e Silva, E. M. d. P. (2013), 'University–

industry collaboration and innovation in emergent and mature industries in new industrialized countries', *Research Policy* **42**(2), 443–453.

Geuna, A. & Rossi, F. (2011), 'Changes to university IPR regulations in Europe and the impact on academic patenting', *Research Policy* **40**(8), 1068–1076.

Griffiths, T. L. & Steyvers, M. (2004), 'Finding scientific topics.', *Proceedings of the National Academy of Sciences of the United States of America* pp. 5228–35.

Grimpe, C. & Hussinger, K. (2013), 'Formal and informal knowledge and technology transfer from academia to industry: Complementarity effects and innovation performance', *Industry and innovation* **20**(8), 683–700.

Indurkhya, N. & Damerau, F. J. (2010), Chapman and Hall/CRC.

Jensen, R. A., Thursby, J. G. & Thursby, M. C. (2003), 'Disclosure and licensing of university inventions:'the best we can do with the s** t we get to work with", *International Journal of Industrial Organization* **21**(9), 1271–1300.

Link, A. N., Siegel, D. S. & Bozeman, B. (2007), 'An empirical analysis of the propensity of academics to engage in informal university technology transfer', *Industrial and Corporate Change* **16**(4), 641–655.

Lissoni, F., Lotz, P., Schovsbo, J. & Treccani, A. (2009), 'Academic patenting and the professor's privilege: evidence on Denmark from the KEINS database', *Science and Public Policy* **36**(8), 595–607.

Munari, F., Rasmussen, E., Toschi, L. & Villani, E. (2016), 'Determinants of the university technology transfer policy-mix: a cross-national analysis of gap-funding instruments', *Journal of Technology Transfer* **41**(6), 1377–1405.

Nomaler, Ö. & Verspagen, B. (2008), 'Knowledge flows, patent citations and the impact of science on technology', *Economic Systems Research* **20**(4), 339–366.

Paukkeri, M.-S. & Honkela, T. (2010), 'Likey: unsupervised language-independent keyphrase extraction', pp. 162–165.

Rong, X. (2014), 'word2vec parameter learning explained', *arXiv preprint arXiv:1411.2738* .

Rothaermel, F. T., Agung, S. D. & Jiang, L. (2007), 'University entrepreneurship: A taxonomy of the literature', *Industrial and Corporate Change* **16**(4), 691–791.

Salter, A. J. & Martin, B. R. (2001), 'The economic benefits of publicly funded basic research: a critical review', *Research policy* **30**(3), 509–532.

Teixeira, A. & Silva, J. M. (2013), 'The intellectual and scientific basis of science, technology and innovation research', *Innovation: The European Journal of Social Science Research* **26**(4), 472–490.

Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D. & Lu, J. (2016), 'Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research', *Technological Forecasting and Social Change* **105**, 179–191.

E CONFERENCE: STI 2018

Title: *Understanding The Diversity Of University Research Knowledge Structures And Their Development Over Time*
Author: Woltmann, Sabrina and Piccolo, Sebastiano and Kreye, Melanie
Conference: 23rd International Conference on Science and Technology Indicators (STI 2018)
Year:*2018*

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Understanding The Diversity Of University Research Knowledge Structures And Their Development Over Time

Sabrina Woltmann[*], Sebastiano A. Piccolo[**], Melanie Kreye[***]

[*]*swol@dtu.dk;*
Management Engineering, Technical University of Denmark, Centrifugevej 372, Lyngby, 2800 (Denmark)
[**]*sebpi@dtu.dk;*
[***] *mkreye@dtu.dk;*
Management Engineering, Technical University of Denmark, Produktionstorvet 424, Lyngby, 2800 (Denmark)

Introduction

Public research in universities is today under high pressure to contribute to society and economic development (D'Este & Patel 2007, Tijssen et al. 2009). Universities are seen as knowledge centres, which means they create new knowledge (Ankrah et al. 2013, Perkmann et al. 2013), provide expertise, and foster innovation (Etzkowitz & Leydesdorff 1997). Universities are knowledge centres and provide expertise, solutions or innovations and inventions (Etzkowitz & Leydesdorff 1997). Accordingly, a key function of universities is knowledge dissemination through different research output types, such as (journal) publications, patents, newspaper articles and so on. This dissemination is often measured through various proxy indicators. Two main approaches can be distinguished: one focusing on research output from academics for academics, such as (journal) publications (Tijssen et al. 2002, Waltman 2016), and the other investigating research output that fosters university-industry exchange, including patents, license agreements and spin-outs (Drucker & Goldstein 2007). However, current methods and empirical studies often focus only on academic or non-academic implications. This separation leads to the absence of recognition of the inter-relation between the different types of research output, resulting in an underassessment of the true impacts of research (Cohen et al. 2002).

This study explores the different types of research output by examining the overall structure of research output of one technical university in Europe over time. The goal is to identify the internal development, relevant key features and their integration into the university knowledge structure (Jensen et al. 2003, Geuna & Muscio 2009). By investigating the structure and changes over time, this study identifies the different dissemination strategies in light of changing paradigms. Our objectives are to investigate the distribution of different output types, to identify their potential content overlap and understand the relevance of these different types. To achieve the objectives we utilize tools from social network analysis and bibliometrics.

Literature

Current studies try to unveil the underlying structures of knowledge transfer from and between universities. This led to highly interdisciplinary research (Gherardini & Nucciotti 2017), focusing either on economic and societal implications (Drucker & Goldstein 2007, Cheah 2016) or on a purely academic perspective (Tartari et al. 2014). The former focuses on

commercially relevant indicators like patents or license agreements (Erdi et al. 2013), while the latter examines academic transfer through citation networks. There has been limited attempt to investigate their relationship (Salter et al. 2017). A recent development is the introduction of 'patent-paper pairs', which uses empirically the combination of patents and their related academic publications (Magerman et al. 2015, Roach & Cohen 2013). For our purpose we draw from the two streams to get a full picture of knowledge structures within one institution. This approach highlights the overall relevance of university research output types. We expect the following outcomes:

> ***Hypothesis 1a:*** There is an observable change in the distribution of the different output types produced by the university over time.
>
> ***Hypothesis 1b:*** Non-journal output becomes more integrated into the network over time.

Furthermore, it is important to identify the overlaps in knowledge between the different types to show the importance of a combined assessment.

> ***Hypothesis 1c:*** Patent-Paper Pairs differ, but overlap, in their references and are bridges to the different partitions in the knowledge network.

Data & Method

This research utilises a network analytic approach because of its suitability for the purpose of this study. Many network analytic approaches are used to grasp the structures and development of knowledge, identifying linkages and emerging topics in various scientific areas (Su & Lee 2010, Zhang et al. 2012, Zhu et al. 2015).

Our sample of research output is collected from one technical university, which has the explicit aim to foster knowledge transfer. We utilised university's own publication database (ORBIT), where all university written output is registered. Our sample contains only entries from the years 2005-2015, since this is the period with most complete data. All entries in ORBIT are registered with a type label, which enables us to distinguish between the different output types like patents, papers, book chapters and a label for the scientific fields (in our case these are classified into 20 different scientific fields). The total number of entries for this period is 77920. We start out with a common citation network created from the Scopus publication database (Boyack 2015, Kamdem et al. 2017), which we generated based on the registered entries from ORBIT. We identify the documents by using string matching for all tiles available. To follow our objectives we add the other types of research output and expand the knowledge network. However, this expansion is by no means trivial and requires quite some additional data processing.

We later add the commercially relevant indicators: patents and their citations, additional open access papers and newspaper articles using additional full-text publications and reference lists. To include these items we need to develop for each new type ways to computationally identify their citations and references. With regard to patents we examine whether these use also internal (university publications) or only external knowledge sources.

Internal Network & External Network

We build an *internal citation network* using only the entries from the university and the links between them. Crucial hereby is to incorporate most available output types and their citations. The identification has to be exercised by another title string matching via the Scopus application programming interface (API). This works satisfactory, in particular for longer titles.

We could identify 28.734 entries from the orbit database in Scopus. These matched entries build the nodes of the internal publication network. Further, we identified in the university database more than 1500 patent applications and retrieved their non-patent literature (NPL). This structure allows capturing the most important and interdisciplinary entries (within the university) of the internal network. On the basis of this internal network we generate also an *external citation network* based on additional Scopus references, which are not output of the university. These are used as measures of external relevance of the publications. This is to assess whether the network structure within the university reflects also the global importance of specific output.

The NPL of the patents shall be used as outward edges, but we also aim to include the patent citations, which show the importance of the inventions. We also aim to investigate the overlap between commercialized and non-commercialized output types of the university research. However, some of the citation identification approaches need improvement. For patents in particular, the integration has not yet been reliable.

Preliminary results

The preliminary results for this study are based solely on calculations that are applied to the basic internal and external Scopus networks. This provides first insights into features of relevant and high quality research items, since these are typically present in the Scopus database. Furthermore, the citations and references are verified and comparatively complete. The overall ratio between registered entries in ORBIT and Scopus is around 40%. The yearly distribution between 2005 and 2015 is not uniform (see Table 1.).

Year	Total university items	Internal network nodes/ external nodes*	Internal network: In edges/ aver. node degree	Internal network: Out edges/ aver. node degree	External network: In edges	External network: Out edges/ aver. node degree
2005	5907	1717 / 48435	4548 / 2.65	301/0.18	62053	44106 / 0.91
2006	6236	1881 / 55408	4836 / 2.57	1025/0.54	67433	50834 / 0.92
2007	6775	2179 / 68047	5414 / 2.48	1767/0.81	76381	62917 / 0.92
2008	6650	2187 / 70074	5319 / 2.43	2527/1.16	76431	65036 / 0.93
2009	6986	2465 / 79742	5740 / 2.33	3410/1.38	75907	74437 / 0.93
2010	6830	2615 / 87398	5729 / 2.19	4429/1.69	74913	82132 / 0.94
2011	7185	3008 / 102628	6412 / 2.13	6159/2.05	78194	97278 / 0.95
2012	7244	2957 / 97430	4588 / 1.55	6150/2.08	54832	93351 / 0.96
2013	7439	3144 / 110493	3687 / 1.17	7103/2.26	50809	107382 / 0.97
2014	7391	3239 / 113894	2275 / 0.70	7690/2.37	42749	112212 / 0.99
2015	7459	3342 / 126416	743 / 0.22	8730/2.61	30950	126391 / 1.00

Table 1: ORBIT papers registered in Scopus per year

** External network nodes have edges with university nodes from the actual year, but no year filtering is applied on the external network nodes.*

In our case, the use of established basic calculations help to identify structural changes.
To compare the networks we apply first simple measures like the average node degree, meaning the average number of links (edges) that a node has. We also distinguish between inwards links (in edges) and outwards links (out edges) generating a directed network.
All nodes, including the university entries that were not found in Scopus, build a large sparse network with 661.859 nodes. Here over 47.000 single nodes have no (identified) connections (the average node degree is then 1.41). Due to this sparsity we remove all unattached nodes. The total number of all remaining nodes is 614.372 with 934.034 edges (1,52 average node degree). The total amount of identified nodes from the university in Scopus from 2005-2015 is 28.734 with 49.291 edges between them (1,72 average node degree).
We examine the development of the network over time by taking snapshots of the different years, calculating specific network properties and compare them. The yearly average in-degree of the internal network show a decrease in the last few years, which makes sense since it takes time before newer publications get cited by new research. The out-degree shows pretty much the opposite trend with a more steady increase in the final years, meaning that the university keeps on using their previous work (see Table 1.). The development of the external network shows similar trends.

An insight provided by the Scopus database is the actual in-edges of each paper. We did not retrieve a full external network and considered only out-degrees from the university entries, but took the overall importance of the papers into account by using their citation scores (Figure 1).

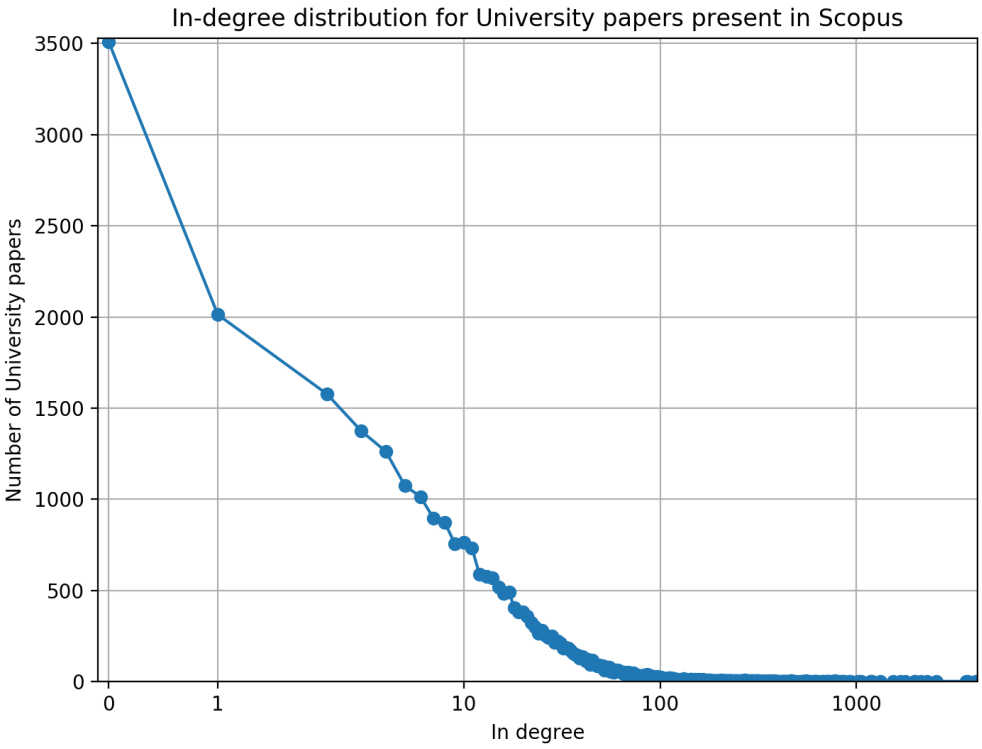


Figure 1: In-degree for University papers present in Scopus

We investigated the changes within the different fields and publication types, like for instance for Open Access. Approximately 25% of university publications in Scopus are Open Access (7192 out of 28734). We looked at the citation count, differentiating for instance Open Access and non-Open Access papers as different types of publications (Figure 2).

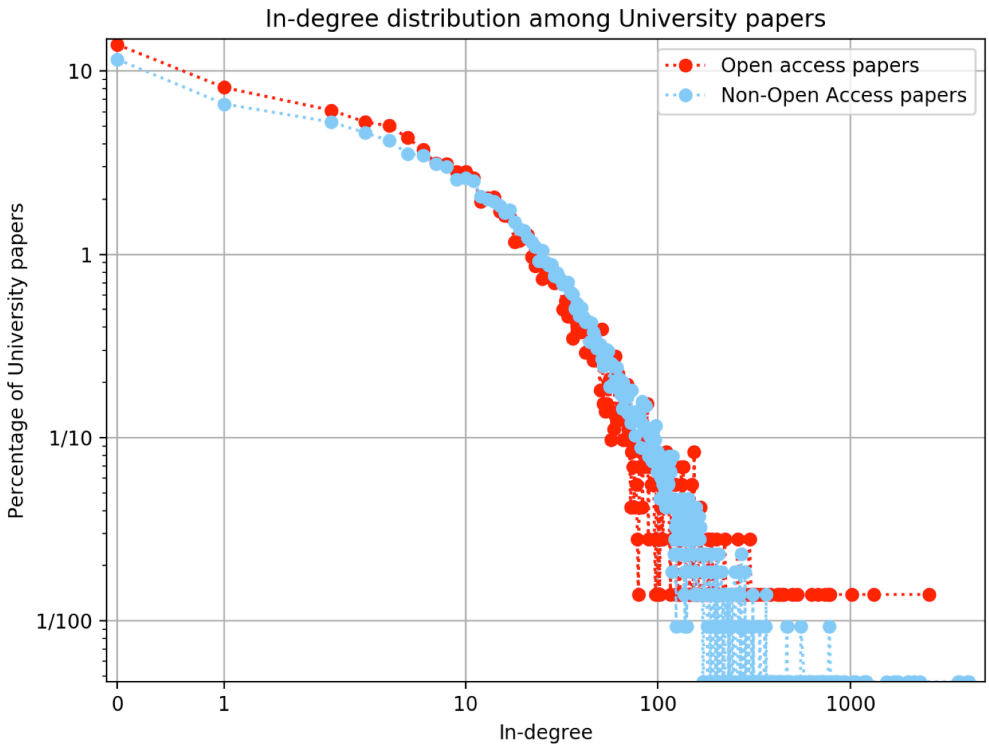


Figure 2: In-degree for University papers present in Scopus based on access type

In the external network Open Access papers do not seem to be more cited, while in fact, it seems that the average non-Open Access publications is usually more often cited. This difference between Open Access and Non-Open Access tends to disappear with highly cited papers (network hubs). When looking at the internal network only, we see a different picture. Table 2 shows the in-degree node ratios. Here, Open Access papers are more central. The average is lower for Open Access due to the low score in the last 2 years and the significant increase in the number of nodes.
Thanks to the comparatively small size of the networks, displaying only one university, a more in-depth insight into network changes is possible. We can see that the total number of open access publications increases from 2011, this is a change as stated in hypothesis 1a) as it shows a clear change in importance of certain output types.

Year	**Open Access Nodes**			**Non-Open Access Nodes**		
	Number of nodes	In-edges	Average in-degree/node	Number of nodes	In-edges	Average in-degree/node
2005	203	551	**2.71**	1514	3997	2.64
2006	249	582	2.34	1632	4254	**2.61**
2007	308	846	**2.75**	1871	4568	2.44
2008	373	1137	**3.05**	1814	4182	2.31
2009	583	1381	**2.37**	1882	4359	2.32
2010	480	1177	**2.45**	2135	4552	2.13
2011	751	2139	**2.85**	2257	4273	1.89
2012	851	1385	**1.63**	2106	3203	1.52
2013	966	1352	**1.40**	2178	2335	1.07
2014	1051	784	**0.75**	2188	1491	0.68
2015	1377	320	**0.23**	1965	423	0.22
2005-2015	7192	11654	**1.62**	21542	37637	**1.75**

Table 2: Open Access vs. Non-Open Access paper in-degrees

Current Challenges
Current challenges are mainly the improvement of title detection in the different data sets.
The data sample has the clear advantage that we are only searching for a limited amount of publications and do not have to rely on the detection of all references in general, which would be even more challenging. However, each of the types has own challenges, which need to be addressed. In particular the detection of citations in the full-texts remains difficult for short titles leading potentially to an under representations of the actual citations.

Discussion

Although we need more research to investigate hypotheses H1b and H1c, we found a difference in trends between open access and non-open access papers, in the internal network. Since 2011, the number of non-open access papers has not been growing, while the number of open access publications has been growing steadily, so we can already state the importance of the internal composition of different output types. The increase of average node degree over years shows an increased importance of the university research within the university itself. This is particularly evident, since the older items have an advantage to be cited also in the following years.

This shows interesting tendencies, but certainly need additional integration of the non-traditional output types into established network, which remains challenging. However, the numbers suggests that this might be highly beneficial. Conceptually, this approach aims to combine the notion of academic and industry knowledge transfer into a combined way of assessing both at the same time.

References

Ankrah, S. N. & Burgess, T. F., Grimshaw, P. & Shaw, N. E. (2013), 'Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit', *Technovation,* 33, 50–65.

Barabási, A. L. (2016), 'Network science', *Cambridge University Press*.

Boyack, K. W. (2015), 'Locating an astronomy and astrophysics publication set in a map of the full Scopus database', *ISSI*.

Buckland, M. & Gey, F. (1994), 'The relationship between recall and precision', *Journal of the American Society for Information Science*, 45, 12-19.

Cheah, S. (2016), 'Framework for measuring research and innovation impact', *Innovation,* 18, 212–232.

Cohen, W. M., Nelson, R. R. & Walsh, J. P. (2002), 'Links and impacts: the influence of public research on industrial R&D', *Management science* 48, 1– 23.

Crespi, G., D'Este, P., Fontana, R. & Geuna, A. (2011), 'The impact of academic patenting on university research and its transfer', *Research Policy,* 40, 55– 68.

Drucker, J. & Goldstein, H. (2007), 'Assessing the regional economic development impacts of universities: A review of current approaches', *International regional science review* 30, 20–46.

D'Este, P. & Patel, P. (2007), 'University–industry linkages in the uk: What are the factors underlying the variety of interactions with industry?', *Research Policy,* 36, 1295–1313.

E ́rdi, P., Makovi, K., Somogyv ́ari, Z., Strandburg, K., Tobochnik, J., Volf, P. & Zal ́anyi, L. (2013), 'Prediction of emerging technologies based on analysis of the us

patent citation network', *Scientometrics* 95(1), 225–242.

Etzkowitz, H. & Leydesdorff, L. (1997), 'Introduction to special issue on science policy dimensions of the triple helix of university-industry-government relations'.

Geuna, A. & Muscio, A. (2009), 'The governance of university knowledge transfer: A critical review of the literature', *Minerva,* 47, 93–114.

Gherardini, A. & Nucciotti, A. (2017), 'Yesterday's giants and invisible colleges of today. a study on the 'knowledge transfer'scientific domain', *Scientometrics* 112, 255–271.

Jensen, R. A., Thursby, J. G. & Thursby, M. C. (2003), The disclosure and licensing of university inventions, *Technical report, National Bureau of Economic Research*

Kamdem, J. P., Fidelis, K. R., Nunes, R. G., Araujo, I. F., Elekofehinti, O. O., da Cunha, F. A., de Menezes, I. R., Pinheiro, A. P., Duarte, A. E. & Barros, L. M. (2017), 'Comparative research performance of top universities from the northeastern brazil on three pharmacological disciplines as seen in Scopus database', *Journal of Taibah University Medical Sciences,* 12, 483–491

Magerman, T., Van Looy, B. & Debackere, K. (2015), 'Does involvement in patenting jeopardize one's academic footprint? an analysis of patent-paper pairs in biotechnology', *Research Policy,* 44, 1702–1713.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Brostr¨om, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A. et al. (2013), 'Academic engagement and commercialisation: A review of the literature on university– industry relations', *Research Policy*, 42, 423–442.

Roach, M. & Cohen, W. M. (2013), 'Lens or prism? patent citations as a measure of knowledge flows from public research', *Management Science,* 59, 504–525.

Schwarz, A. W., Schwarz, S. & Tijssen, R. J. (1998), 'Research and research impact of a technical university—a bibliometric study', *Scientometrics,* 41, 371–388.

Su, H.-N. & Lee, P.-C. (2010), 'Mapping knowledge structure by keyword cooccurrence: a first look at journal papers in technology foresight', *Scientometrics* 85, 65–79.

Tartari, V., Perkmann, M. & Salter, A. (2014), 'In good company: The influence of peers on industry engagement by academic scientists', *Research Policy* 43, 1189–1203.

Tijssen, R. J., Van Leeuwen, T. N. & Van Wijk, E. (2009), 'Benchmarking university-industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities', *Research Evaluation,* 18, 13–24.

Tijssen, R. J., Visser, M. S. & Van Leeuwen, T. N. (2002), 'Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?', *Scientometrics* 54, 381–397.

Waltman, L. (2016), 'A review of the literature on citation impact indicators', *Journal of Informetrics* ,10, 365–391.

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. & Lu, Z. (2012), 'Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis', *PloS one,* 7

Zhu, L., Liu, X., He, S., Shi, J. & Pang, M. (2015), 'Keywords co-occurrence mapping knowledge domain research base on the theory of big data in oil and gas industry', *Scientometrics,* 105, 249–260.

F CONFERENCE: EUSPRI 2018

Title: *"Do Open Access Publications Facilitate University-Industry Knowledge Transfer? - A Novel Perspective -"*
Author: Woltmann, Sabrina and Alkærsig, Lars and Lomberg, Carina
Conference: Eu-SPRI Conference Paris Year: *2018*

Do Open Access Publications Facilitate University-Industry Knowledge Transfer? - A Novel Perspective -

Introduction & Motivation

This study investigates the difference between university research output in subscription based and freely available open access publications. The study focuses in particular on the differences in knowledge transfer to the industry of those two types of publications. We apply computational linguistic methods to investigate the differences in success regarding the dissemination of university research output.

Introduction

University research is seen as a main driver for innovation, contributing both to firm and societal economic growth (Ankrah et al. 2013, Perkmann et al. 2013). As part of their role in society, universities often function as external knowledge sources, providing external impulses companies need to progress in their R&D (D'Este & Patel 2007, Tijssen et al. 2009, Cohen et al. 2002). Universities are evaluated by their 'contribution', as an external knowledge source the industry can derive from (Janeiro et al. 2013).

Clearly, open access publications are seen as a way to ensure public access to research outcomes of the university. Hence, the effect of open access is a relevant question for knowledge dissemination (university perspective) accessibility (industry/ society perspective). Most of the studies have focused on the impact of open access publications for the scientific community in terms of improvement of citation count for the researcher(Antelman 2004). This literature is spread over different types of journals and scientific fields ranging from bio-medical science and ecology to telecommunication science(Tang et al. 2017). Most fields focus either on the development in their field of Open Access and its impact on scholarly citations, or on the journal landscape (Wang et al. 2015, McKiernan et al. 2016). A certain reservation regarding open access publications is also evident in this rather young field focusing mainly on the open access journals that exploit the author-pays model, which are often seen to damage scholarly publishing (Beall 2012). Some argue that the rise of open access journals and

publications might harm the research integrity, due to the increasing amount of Open Access non- peer reviewed journals. At the same time giving new impulses and lowering publication time and costs. But is this knowledge accessible and if so for whom? In recent years, research on open innovation and all related issues became much more fundamental, ethical and developmental issues and consequently lead to free accessibility of publicly funded research outcomes becoming a major keystone in many new policies. The United Kingdom and other countries incentive open access publishing [1]. The expectations of policy makers regarding the potential impact and improvements seem to be rather high, but seemingly without a sound scientific basis. Despite the recent research efforts, there has been little focus on a comparison between open access publications and subscription based publications, which are not based on bibliometric methods and the anticipated positive effect on the university-industry knowledge exchange remains to be an intuitive, but never the less not fully understood and only limited evidence based.

We see the need to introduce this matter to the community that is concerned with university-industry collaboration and public research impact and studies of only the scientific community itself. It is time to investigate another very relevant level of open access publications: The knowledge dissemination and knowledge transfer effects.

Overall, it is crucial to understand this aspect of open access publications, not only for practical reasons of scholars and policy makers, but also to enhance the theoretical foundation of knowledge dissemination. If the impact of university research is discussed in today's society this aspect needs to be part of the scholarly debate. We need to know more in detail how much impact it has if universities publish open access, of course the idea is availability of knowledge, but it could also not have the desired effect. Little is known about the usage differences of open access via subscription based articles and there is a need to shed light into this matter.

The objective of this study is to use statistical learning and text mining methods to trace knowledge transfer from university to industry, while at the same time distinguishing between open access publications and subscription based articles.

Method

Understanding the impact of the difference in the dissemination success rates for open access and subscription based university publications is a challenging task, which this study addresses by using techniques from the domain of computational linguistics (Rus et al. 2013, Crockett et al. 2017).

We verify the dissemination rates of both types by using these techniques to trace research content from publications and identify the same content in related firm documents (obtained from their websites). This coherent contents are identified via an combination of established statistical methods from the field

[1] url http://www.hefce.ac.uk/rsrch/oa/Policy/

of natural language processing (NLP). The methods from NLP help to manage large amounts of text data, reduce their dimensionality to enable more efficient pattern recognition. However the state-of-the-art in text similarity measures is still developing and not quite in a point where the methods are advanced enough to provide an effective pre-made solution for our problem. Therefore we need to use a combination of them to fulfill this complex task. For data preparation, cleaning and adjustment we follow the common standards of the community including pre-processing, word stemming, and language recognition (Indurkhya & Damerau 2010).

To understand our samples and results better we use the commonly used topic model latend dirichlet allocation (LDA) to identify the main topics each of the companies or respectively research areas are engaged in (Blei et al. 2003). These generated topics are used to identify the general area a text is related to. We combine this method with a simplistic algebraic method, Term-frequency inverse document frequency (TFIDF), which is generally applied to reflect how important a single word is to a documents content. This is used to extract relevant keywords of the documents and to allow word co-occurrence comparisons[2]. We combine findings out of this method with an more advanced method, the latent semantic analysis (LSA), which is based on single value decomposition (SVD) and applies a specific form of rank lowering to reduce the dimensionality of the data (Landauer et al. 1998, Niraula et al. 2013). It allows to grasp the relationships between a corpus of documents and the terms the documents contain, therefore it also allows the comparison of different documents. This method is comparatively old, but performs good on smaller data sets (especially in comparison to deep learning and neural network algorithms that need an immense amount of data to perform satisfactory).

These methods, if combined, can find statistically similar contents in text documents. For similarity measures we use the Jaccard coefficient (TFIDF) (Zhang et al. 2016) and the cosine similarity (LSA). However, since the development of text similarity is still a not solved computational problem, we need human verification to confirm the computer identified matching text pairs.

The true positives (academic texts matching company texts) are then in turn identified as open access or subscription based articles. In the final step we investigate the differences between the open access publications and subscription based publications that have their content in company texts. We account for research area, time the paper was published, and the citation score [3]

[2]For more detailed information on the technical specifications see (Woltmann & Alkærsig 2017)

[3] This is needed to be able to put the paper into its context, meaning the relevance it has for the scientific community. If we would not take it into account we might have only highly cited papers in one type which limits the potential to generalize the findings.

Case & Sample

This study uses the case of Technical University of Denmark (DTU). This university is an ideal match since it is dominated by natural and technical sciences being a technical university. In particular since it has a strong focus on applied sciences and is also Denmark's university with most industry collaborations, hence highly relevant for economic and technical development in its environment. Furthermore, DTU's data availability and accessibility in terms of publication records, open access as well as subscription based, and collaboration partners, made DTU a good fit for a first assessment.

Publications: Open Access & Subscription Based

The first text document sample contains all available academic publications registered at the university publication database (ORBIT) between 2005-2015. For most of these publications an abstract text is available, which is used as text document for the statistical analysis. The total number of publications is 77920 with 23818 of those identified as open access publications and 54102 as subscription based articles. These are classified by us into 20 different scientific fields of university research, including very novel ones and comparatively small ones, such as Nanotechnology, Biosystems, Photonics, Space and more traditional ones, like Mechanical Engineering, Civil Engineering and more. However, we have to take into consideration that the amount of open access publications increased highly during the last years, which might effect the findings.

Company text documents

The sample of company related documents is performed by crawling websites[4] of a sample of companies accessing web-page texts and PDF documents available on the firm's website. As a starting point we chose to explore websites, which belong to university collaborators, meaning that they had a contract with the university between 2006 and 2016. These were obtained by the legal department of the university, where collaborations and contracts are registered. We obtained a sample of around 1200 different company names [5].We additionally used the universities own website to identify further collaborators via hyperlinks. To limit the sample to a manageable size we introduced additional criteria for the companies (these criteria are identified in the raw text data of the company websites and rely on detection and correctness of the company material): First, a Danish registration number (CVR) or an *'ApS'*[6] in the name. Second, English text documents on the website with at least 5 web-pages and more than 100 words for the whole website. After the selection we are left with a final sample of 454 company websites with usable text documents.

[4]For more technical insights please see the GitHub repository of the WebExplorer
[5]Small bilateral agreements between professors and companies, which do not require approval might not in all cases be registered. However, we did not see any evidence for missing data
[6]Describing a private limited company 'Anpartsselskab' in Danish.

(Preliminary) Results

The application of the LDA shows clearly that the topics of certain industry partners of the university are very aligned to the universities scientific areas. However, it seems that particular fields are over represented, which make a positive detection in those scientific fields of course more likely, but reflect the importance of the fields. We plan to compare these findings with the ratios of the open access and subscription based publications, to be able to ensure that the distribution is even, and in case it is not to account for it in our analysis.

The statistical analysis based on the keyword extraction with the TFIDF using word co-occurrence measures shows that more than 50% of the identified positive matches between company websites and publications were open access publications. This might have several potential explanations such as: a) the open access publications are in more 'interesting' or 'relevant' domains for companies, b) the open access journals use more terms which are also used in the industry, c) they publish in areas with many very succinct terms or proper nouns, d) are just newer and more on the cutting edge.

However, this is a high number especially considering that the entire sample only has 30% open access publications. Overall we found most statistical matches in the fields of:

- Electrical Engineering
- Environmental Engineering
- Food
- Mechanical Engineering
- Photonics (most positive matches)
- Space

Combining the results from the LDA with the once from the TFIDF (by combining keyword lists and topics in different ways) did not yet lead to an improvement of the results, but we are confident to reach improvements if these are combined with the LSA outcome.

The LSA has in the first investigation shown (applied only on publications from the scientific fields of Photonics, Space Research and Nuclear Science) that the similarity between the websites and academic corpora is very low, which is a great indicator, meaning that really rare high matching score are likely related to true positive findings. Additionally the documents identified as most similar (statistically) vary from the ones the TFIDF returned, meaning these both methods capture clearly different and potentially complementary things.

However, the manual verification for this process is still ongoing but current interim status show that here again the open access publications are dominating the positive sample. This has to be taken with caution, since the true positive rates are very small and not yet fully evaluated. Due to the highly specialized

literature of the academic sample we need several highly educated validators to be sure about the reliability of the judgments.

In addition to the used abstracts we obtained full-text publications (open access and subscription based publications) to verify our results in the future. However, abstracts are widely used in empirical biblometric and text analyses, but we assume that full-texts might outperform them significantly.

Discussion & Conclusion

Preliminary results suggest that the open access publications in our sample actually contain more research knowledge that is used by companies.These are only preliminary results that have to be expanded. Additionally we aim to identify all the potential variables that might explain the outcome in favour of the open access papers, which are not related to the access status of the paper. Hence we are currently working to address the potential influence of other variables in our sample by investigating the time and field distributions and relevance of the single papers. However, since this is a novel approach there is not much empirical work done prior to this study where we could draw from.

From a policy perspective this different approach offers, even if not yet fully established, an additional insight into the potential relevance of open access publishing. This new perspective is very important, as it focuses on the impact of open access on the usage in industry. Knowing the impact open access has on the academic community is certainly highly relevant, but we believe that the impacts on this particular community might not hold up for other contexts. Hence, practitioners as well as academics need to seek ways to understand and measure its impact in societal and economic dimensions.

It is important to expand our knowledge about the impact of freely available research driven knowledge to gain a better understanding of it. Its positive and potentially negative impacts have to be assessed to ensure that public research keeps promoting innovation and development in the most effective manner. We believe our study contributes to this with a novel perspective on the relevance of open access publications.

We aim to answer structural questions, such as whether the outcome structure changes, looking in particular at the increase/decrease of patents, newspaper articles and academic papers over time. It is crucial to understand how the different types of outcomes are connected and whether some might contribute or draw from commercialization processes.

References

Ankrah, S. N., Burgess, T. F., Grimshaw, P. & Shaw, N. E. (2013), 'Asking both university and industry actors about their engagement in knowledge transfer: What single-group studies of motives omit', *Technovation* **33**(2-3), 50–65.

Antelman, K. (2004), 'Do open-access articles have a greater research impact?', *College & research libraries* **65**(5), 372–382.

Beall, J. (2012), 'Predatory publishers are corrupting open access: journals that exploit the author-pays model damage scholarly publishing and promote unethical behaviour by scientists, argues jeffrey beall', *Nature* **489**(7415), 179–180.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Cohen, W. M., Nelson, R. R. & Walsh, J. P. (2002), 'Links and impacts: the influence of public research on industrial r&d', *Management science* **48**(1), 1–23.

Crockett, K., Adel, N., O'Shea, J., Crispin, A., Chandran, D. & Carvalho, J. P. (2017), Application of fuzzy semantic similarity measures to event detection within tweets, *in* 'Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on', IEEE, pp. 1–7.

D'Este, P. & Patel, P. (2007), 'University–industry linkages in the uk: What are the factors underlying the variety of interactions with industry?', *Research policy* **36**(9), 1295–1313.

Indurkhya, N. & Damerau, F. J. (2010), *Handbook of natural language processing*, Vol. 2, CRC Press.

Janeiro, P., Proença, I. & da Conceição Gonçalves, V. (2013), 'Open innovation: Factors explaining universities as service firm innovation sources', *Journal of Business Research* **66**(10), 2017–2023.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998), 'An introduction to latent semantic analysis', *Discourse processes* **25**(2-3), 259–284.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K. et al. (2016), 'How open science helps researchers succeed', *Elife* **5**.

Niraula, N., Banjade, R., Ştefănescu, D. & Rus, V. (2013), Experiments with semantic similarity measures based on lda and lsa, *in* 'International Conference on Statistical Language and Speech Processing', Springer, pp. 188–199.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A. et al. (2013), 'Academic engagement and commercialisation: A review of the literature on university–industry relations', *Research policy* **42**(2), 423–442.

Rus, V., Niraula, N. & Banjade, R. (2013), Similarity measures based on latent dirichlet allocation, *in* 'International Conference on Intelligent Text Processing and Computational Linguistics', Springer, pp. 459–470.

Tang, M., Bever, J. D. & Yu, F.-H. (2017), 'Open access increases citations of papers in ecology', *Ecosphere* **8**(7).

Tijssen, R. J., Van Leeuwen, T. N. & Van Wijk, E. (2009), 'Benchmarking university-industry research cooperation worldwide: performance measurements and indicators based on co-authorship data for the world's largest universities', *Research Evaluation* **18**(1), 13–24.

Wang, X., Liu, C., Mao, W. & Fang, Z. (2015), 'The open access advantage considering citation, article usage and social media attention', *Scientometrics* **103**(2), 555–564.

Woltmann, S. & Alkærsig, L. (2017), Tracing knowledge transfer from universities to industry: A text mining approach, *in* 'Academy of Management Proceedings 2017 (AOM)'.

Zhang, Y., Shang, L., Huang, L., Porter, A. L., Zhang, G., Lu, J. & Zhu, D. (2016), 'A hybrid similarity measure method for patent portfolio analysis', *Journal of Informetrics* **10**(4), 1108–1130.

G CONFERENCE: DRUID PHD ACADEMY 2018

Title: *Trace of Knowledge: Benchmarking Novel Text Mining Based Measurements*
Author: Woltmann, Sabrina
Conference: DRUID PhD Academy Conference 2018
Place: Odense, Denmark Year: *2018*

er to be presented at the DRUID Academy Conference 2018 at University of Southern Denmark, Odense, Denmark on January 17-19, 2

Trace of Knowledge: Benchmarking Novel Text Mining Based Measurements

Sabrina Larissa Woltmann
Technical University of Denmark
DTU Management Engineering
swol@dtu.dk

Abstract

The impact of public research outcomes on economies, and societies, in particular, in terms of innovation and development is widely accepted and empirically investigated [9, 3]. However, many studies suggest a systematic underestimation of the impact and benefits of public research. Empirical studies describe that current approaches capture only specific aspects of knowledge transfer between public research institutions and private entities [1, 4, 7].

The main interrelated reasons contributing to this systematic underestimation are that most established knowledge transfer measurements focus on intermediaries and use proxy-indicators like patents, licenses, spin-outs and co-publications as data sources, but these metrics are problematic because they can result in type I and type II errors, since many of them capture a transfer that is never utilized by a private entity (e.g. like unused patents) [8]. In addition, there are occasions where the proxy is not met so the actual use is not being captured.

We try to improve this systematic underestimation by adapting novel computer linguistics methods to this field and putting them into perspective with the existing measures of knowledge transfer. We use both basic and more advanced statistical learning tools from the field of computational linguistics and statistical learning to trace the knowledge fragments[2, 6]. In addition, we utilize a mixture of standard algebraic and probabilistic methods. Furthermore, pattern recognition classification algorithms help to trace the public research outcomes, going beyond plain word co-occurrence[10, 5].

We identify concrete public research outcomes that identify their transfer and trans-

Jel Codes: O31, O33

Trace of Knowledge: Benchmarking Novel Text Mining Based Measurements

Sabrina L. Woltmann

Introduction

Public and politics increasingly expect universities to contribute to societal and economic developments, with their publicly funded research. They are expected to provide strategic dissemination of their findings, which expands their traditional tasks, teaching and research, complementing the two prior missions as it translates the research efforts into economic and societal relevant contributions. To conform to these demands, universities implemented various forms of transfer activities (Gulbrandsen & Slipersaeter 2007) to promulgate their research outcomes in order to ensure that their findings are utilized, thereby securing public funding.

Universities are important for the society as producers and transmitters of novel knowledge and technologies and as supporters for the economy to tackle given challenges, or help to improve already existing solutions (D'Este & Patel 2007, Zawdie 2010, Walsh et al. 2016). Ideally, the generated knowledge is received and used by the industry or other relevant third parties and contributes to the development of the socio-economic environment of the university (Agrawal & Henderson 2002). Scholars agree that research driven innovations can lead to economic growth, development and increased competitiveness (Huggins et al. 2008, Vincett 2010). Accordingly, public research outcomes benefit economies and societies, in particular, their influence in terms of innovation and development is widely accepted and empirically investigated (Zawdie 2010, Drucker & Goldstein 2007).

However, many studies suggest a systematic underestimation of the impact and benefits of public research. Empirical studies describe that current approaches capture only some aspects of knowledge transfer between public research institutions and private entities (Agrawal & Henderson 2002, Lissoni 2012, Roach & Cohen 2013). The main interrelated reasons contributing to the systematic underestimation are that most established knowledge transfer measurements focus on intermediaries and use proxy-indicators like patents, licenses, spin-outs and co-publications as data sources. These metrics can result in errors, since many of them capture a transfer that is never utilized by a private entity (e.g. unused patents or co-publications) (Walsh et al.

2016, Cheah 2016, Lundberg et al. 2006). Furthermore, these indicators face long-standing criticism about their incapability to capture the majority of transferred knowledge. The indicators for economic impact of quantitative assessments are *'(. . .) difficult to obtain and generally suffer from long lag times between public investment and outcomes'* (Arundel & Bordoy 2008, p.6). Besides, it has been pointed out that they fail to provide a holistic picture. Some scholars argue that the measurements are only accounting for very low percentages of actual knowledge transfer (Cheah 2016, Lundberg et al. 2006).

Traditionally most commonly used proxy-indicators enabled scholars to trace particular inventions, innovation and (public) knowledge flows. Patents, for instance, describe novel technological solutions to problems, and at the same time reference the prior knowledge used for the invention, including other patents and other scientific and academic literature(Roach & Cohen 2013). These references are often used to identify the flows of knowledge and expertise within technical innovation; they even allow to track the interrelations between different innovations. This approach has provided useful insights over the years, but it does not capture the entire knowledge flow from public research institutions to the industry, due to several limitations: first, it traces only codified explicit knowledge; second, it captures only patentable inventions; third, the data can suffer from selective non-patent referencing; fourth, it is highly industry dependent (some industries just patent less than others); fifth, it does not account for most of the basic research use in the industry; sixth, it does not necessarily trace actual commercial use as a patent might never lead to an actual commercialization.

Due to the above mentioned shortcomings and the lack of supplementary indicators or methods, this approach faces longstanding criticism. Hence, in order to improve the detection and understanding about knowledge flows, additional comparable measures of knowledge flows are needed.

Motivation

Our motivation is to offer additional measurements that can shed light on knowledge flows and the use of public research outcomes to ensure that policy makers and other involved parties have a more coherent overview about dissemination of knowledge and the socioeconomic benefits arising therefrom. Given the limitations of contemporary empirical work, we contribute to the body of academic literature by addressing some of the deficits of the current metrics. Our method provides in-depth insights about the transferred knowledge. Many studies, lack a clear definition of the concept of 'knowledge', as

well as of 'knowledge transfer'(Liyanage et al. 2009). The term itself is highly debated and conceptualized in various philosophical approaches; to limit it to a reasonable scope, we focus on the definition relevant for this particular context. Therefore we use the foundations of knowledge management theories and focus on the explicit knowledge, which *"(...) involves know-how that is transmittable in formal, systematic language and does not require direct experience of the knowledge that is being acquired (...)"* (Howells 2002, p. 872), while we do not include tacit knowledge, *"non-verbalised, intuitive and unarticulated knowledge"* (Polanyi 1962) in our study. Additionally, this study focuses only on research related and novel knowledge. The scope comprises knowledge and technologies, which are potentially relevant for future innovation processes and are novel to the scientific community. This excludes widely known and commonly accepted basic knowledge as it is, for instance taught at universities. Knowledge transfer and technology transfer are in the body of literature extremely interrelated concepts and thus often used in an exchangeable manner (Agrawal 2001, Grimpe & Hussinger 2013, Sung & Gibson 2000). We, however, focus on knowledge transfer overall, but acknowledge that the term 'technology transfer' is, in certain cases, a more accurate description of the issue.

To improve the understanding about knowledge flows from universities to the industry, we aim to expand the insights by using patent descriptions and compare them to knowledge flows that are traceable via text mining methods. It is critical to identify whether different types of innovations are traceable via both methods, or whether a single method performs better for some cases.

Fortunately, texts contain information, and extracting this information has become an increasingly developed part of today's research fields of computational linguistics, which makes these tools available and applicable in other academic disciplines. Insights in various disciplines are generated via the use of text mining tools over the past decades. Especially content analysis are of increasing relevance in machine learning and the tools advancement is getting more and more promising (Chapman & Hall/CRC 2010, Collobert et al. 2011).

Particularly for our case, text mining is an appropriate strategy since text material is a promising data source to detect different parts of knowledge. 1. Academic publications, such as journal articles, conference proceedings or books, contain the main outcomes of scientific research. They are seen as output and at the same time dissemination channel of university research (Stahl et al. 1988, Toutkoushian et al. 2003). Therefore these publications are texts containing data for all major research findings of a university. 2. Patents are seen as key indicators for inventions and describe specific knowledge generated by a certain entity or person, which makes them as well a indicator

for knowledge generation, as well as a transfer channel. 3. Online websites of companies are textualized media for companies to display their novel products, services, R&D strategies and innovations, publishing their actual utilization of knowledge. Companies place high value on these to ensure their visibility for potential consumers and investors leading to regular updates and R&D descriptions (Branstetter 2006, Heinze & Hu 2006).

These types of texts provide insights into the use and generation of knowledge. Therefore, the use of statistical tools from *natural language processing* (NLP) is an potential new approach to identify commonalities and correlation between the data sources.

We will use a combination of different NLP tools to identify pieces of public research chunks and to pinpoint overlaps and discrepancies in the results. We determine which types of innovations lead to observable knowledge transfer.

Strategic Approach

The following is a brief summary of data analysis and the application of the text mining methods.

The academic publications sample, including abstracts and open access texts, was compared to patents, to identify common topics and identical content. This process reveals the inter-relatedness of the two samples,patents and publications, and shows their great content overlap . In addition, this approach has the potential to allow for the use of patent references to give evidence as to whether patent citations display the entire content of *prior art*[1], or whether we may be able to find additional, not mentioned prior art research.

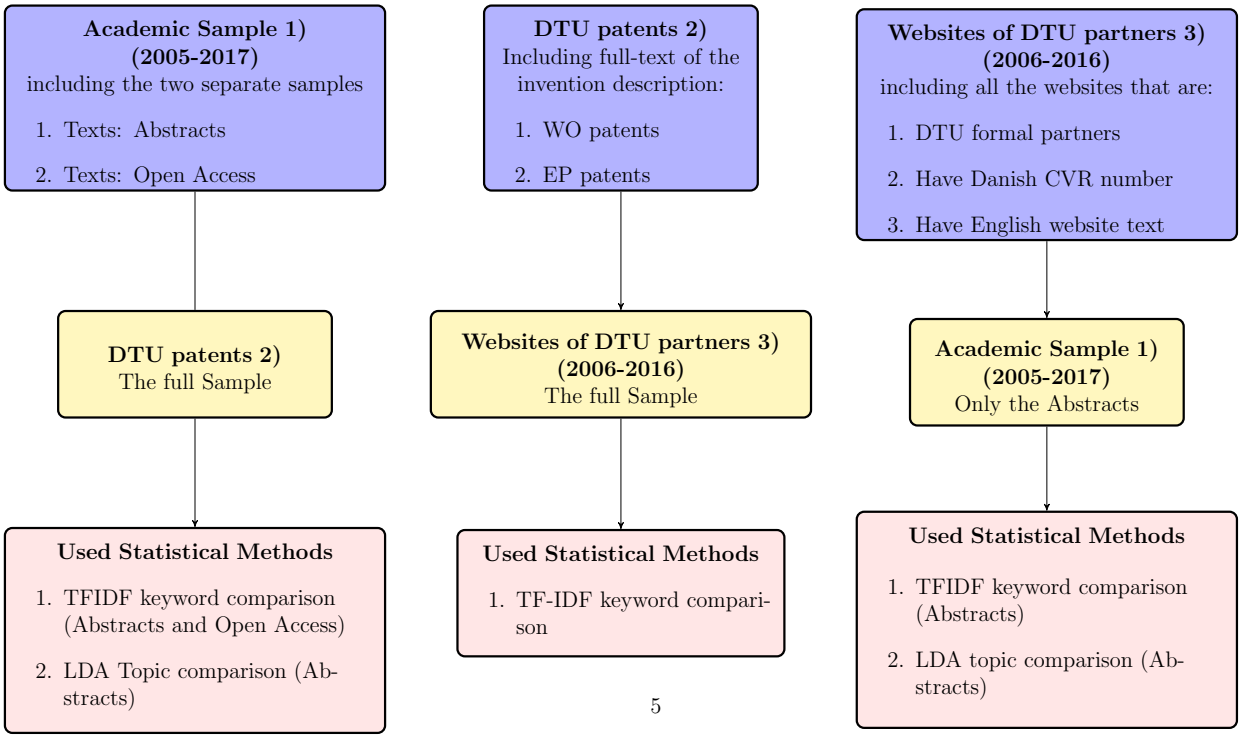
This study uses the academic publications open-access text samples and abstracts, and compares them with company homepage texts. The statistical methods applied in this case were TFIDF as well as the LDA.

The university patents are also compared to the homepage texts to identify minimal and congruent content. These outcomes can be compared to the outcomes we derive from the previous step, and this process can demonstrate whether knowledge: a) that is in patents is more likely to get used by the industry, b) that is in both patents and publications is more likely to be used, or c) that knowledge that is in patents wis less likely to get utilized by industry.

[1]Also called 'state of the art,' which refers to prior relevant published material that supports the understanding of the invention and hence often represents knowledge needed for the invention. However, it has to be kept in mind that many prior art items are included by attorneys or agents for legal reasons, but may not be relevant for the subsequent patent.

Text comparisons to identify Knowledge Transfer

Samples used in different comparisons



Method

The purpose of this study is among others to verify and improve the use of computational linguistic tools and concepts for the detection of knowledge flows. Hence, the study compares texts and text snippets which potentially contain research knowledge with available statistical methods from NLP. We aim to trace chunks of the same knowledge in different texts by using different methods that allow pattern recognition and co-word identification. The data pre-processing and undertaken steps are detailed in the following subsections.

Pre-processing and statistical units

Text pre-processing converts unstructured raw text into computational useful units, which in turn can be used to generate meaningful input for the application of statistical methods. Pre-processing is an important aspect of text analytics procedures and might very well be decisive for the quality of the obtained results. The main objective is to capture all relevant terms and characters and at the same time to identify and erase obsolete terms. Additionally, it is crucial to classify actual words via tokenization and try not to lose too much relevant textual information, while cleaning the texts. All this is important for the application of further text mining methods (Paukkeri & Honkela 2010).

Pre-processing of text included:

- Removal of unwanted elements (e.g. special characters, punctuation and numbers),
- Conversion from upper case to lower case,
- Removal of 'stopwords' [2],
- Stemming of the terms [3].

The pre-processing revealed some challenges in the case of the academic texts. They contain, for instance, chemical formulas and notations, which rely on numbers and/or special characters. These are lost during the pre-processing

[2]Stopwords are the most common words in a language, which are not carrying content relevant information

[3]Describes the process of reducing words to their word stem or root form. It is a process for removing the morphological endings from words: connected, connection, connections become 'connect'. We used the Porter stemmer build in in the CRAN tm package in R `https://cran.r-project.org/web/packages/tm/index.html`

and the only possibility to identify the same formulas for similarity measures is that the removal will always result in the same character string [4]. Short strings derived from formulas or notations are lost during the pre-processing. Certain terms seem to be the result of poor pre-processing, but are in reality just a representation of specific models, pronouns or even project names.

The pre-processed texts are logically grouped into a text collections denoted text corpus, and in our case into several text corpora. All applications of statistical methods are based on these corpora.

The corpora are converted into *term-document matrices*, which is the most common vector space representation of document corpora. It represents the frequency for each term in for each document. Rows correspond to documents and columns to terms.

A document-term matrix is generated from pre-processed corpora and therefore represents contextual relevant terms (Chapman & Hall/CRC 2010). As document-term matrices are for the most part highly dimensional and sparse matrices, most models include some sort of sensible dimensionality reduction (Berry & Castellanos 2007). In a document-term matrix the element at (i,j) is the word count (frequency) of the i'th term (t) in the j'th document (d).

$$TermDocumentMatrix = \begin{array}{c} \\ d_j \end{array} \begin{array}{c} \begin{array}{cccc} & & t_i & \\ \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & ... & x_{1,j} \\ x_{2,1} & x_{2,2} & ... & x_{2,j} \\ \vdots & & \ddots & \\ x_{i,1} & x_{i,2} & ... & x_{i,j} \end{pmatrix} \end{array}$$

Different weighting schemes, determining the value of each term entry, have been developed. These are derived from frequencies of the term occurrences per document. The most relevant weighting scheme to use is highly depending on the statistical models that are used to analyze the data. Common weighting schemes include, a binary weighting, where the entries takes values 1 or 0 depending on whether or not a term occurs; Term-frequency (TF), the actual number of times a term occurring a document and the Term-frequency; inverse document frequency (TFIDF), uses TF but assigns higher weight to terms that occur only in a small number of documents.

[4] In some cases HTML tags prevent the identical re-construction. In this case we did not yet find a way to identify the matching strings.

Term-frequency, inverse document frequency

The TFIDF is a simple numerical indexing method, which has been applied in various contexts (Franceschini et al. 2016, Zhang et al. 2016). It has proven to give respectable results on its own, especially considering its simplicity. It is also used as basis in various models, such as Latent Semantic Analysis (LSA) (Mao & Chu 2007). TFIDF is an indexing scheme that allows identifying the most relevant words. It allows for instance to reduce the dimensionality of texts to a small set of terms, capturing the main content of a document. The principal assumption hereby is that a word that occurs often in a document is relevant for its content, but words that are used in many documents are less specific to q single document. TFIDF has different proposed calculations, most commonly it is calculated by multiplying the term frequency TF, the number of times a word w appears in a document d; and the inverse document frequency IDF, which is the logarithm of the total number of documents D divided by the number of documents that contain the word w denote dw.

$$TF(w,d) = \sum w_i$$

$$IDF(w,D) = log(\frac{D}{dw})$$

$$TFIDF = TF(w,d) \times IDF(w,D)$$

TF calculates term weight based on term frequencies might represent merely the content of a text fragment. The IDF part ensure the representation of the most distinct terms from the other documents in the collection (Xia & Chai 2011).

For the purpose of this study, we chose to use the TFIDF indexing to determine the 50 most characteristic words per document. In case of the academic abstracts we retrieved often less than 50 words. The reduced dimensionality enables a comparison of keywords. We use the keywords, generated for each document, to identify common terms between the different documents.

Latent Dirichlet Allocation

LDA is an unsupervised algorithm that performs *topic modeling*. It is based on statistical learning and aims to identify unobservable underlying structure within a text collection (Blei et al. 2003, Griffiths & Steyvers 2004). LDA extracts these structures and translates them into topics. These topics are composed of terms that are assigned together with a certain probability.

LDA is described by Grün and Hornik with the following steps (2011, p. 4) and Ponweiser(2012, p.15):

1. For each topic: decide what words are likely (term distribution described as $\beta \sim Dirichlet(\delta)$

2. For each document:

   (a) decide what proportions of topics should be in the document, (topic proportions defined by $\theta \sim Dirichlet(\alpha)$.

      i. for each word in the document:

         A. choose a topic ($z_i \sim Multinomial(\theta)$).

         B. given this topic, choose a likely word (generated in step 1.) from a multinomial probability distribution conditioned on the topic $z_i : p(w_i|z_i, \beta)$.

LDA requires the number of topics (K), to be chosen beforehand, hence we chose to approach this in two ways: we either approximate the marginal corpus likelihood (depending on K) by taking the harmonic mean of the corpora after applying the LDA with different number of K, or simply by corpus size assuming that a larger corpus contains more distinct topics:

$$D_m \geq 3000 : K = 200$$

$$D_m \geq 2000 : K = 150$$

$$D_m \geq 1000 : K = 100$$

$$D_m \leq 1000 : K = 50$$

(If applied the harmonic mean we limited the maximum number of topics for the corpora of websites in each case to 200 and found that only 3 web corpora might have benefited from a larger number, we chose this limitation for computational efficiency reasons.) For all the LDA application the hyper-parameter was set so that they enforce more topics per document, but with lower probabilities (Grün & Hornik 2011). This is necessary to capture context and topics of text snippets. To further improve the performance, we excluded terms, which occur in more than 90% of the documents in the document-term matrix. The resulting topics are very specified, especially after the additional pre-processing step, but kept these terms as a separate keyword list.

We extracted the 50 best terms per topic (the ones with the highest probability) and returned them as list of keywords. We compared to other lists of topic keywords from LDAs from academic corpora and web corpora. The resulting topic pairs show the most similar corpora in terms of their underlying structures.

Jaccard Similarity Coefficient

For assessing the similarity between sets of identified keywords found by applying TFIDF or LDA, we used the Jaccard similarity coefficient as metric. It is a statistic used for measuring the similarity between sets. The Jaccard similarity is based on the size of the intersection divided by the size of the union of the sets. The measure is between 0 and 1, 1 indicating most similarity (identical sets) and 0 indicating least similar: no common feature in the two sets. Given the set of keywords from one text denoted K_A and the another set denoted K_B, the Jaccard similarity $J(K_A, K_B)$ is obtained with:

$$J(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|} = \frac{|K_A \cap K_B|}{|K_A|+|K_B|-|K_A \cap K_B|}$$

We chose this similarity measure as it only includes element presence in a set and is therefore applicable without relying on potentially not comparable scores. Hence, it can be used for the LDA and TFIDF. A Jaccard Similarity coefficient threshold needs to be determined for considering both keywords sets to be matching one another. The minimum Jaccard similarity threshold was so far always adapted to the outcome of the texts compared and adjusted to find around the 50 best matches. However, due to the varying size of keywords sets, we defined that pairs of sets with Jaccard Similarity lower than 0.15 needs more than 7 words in common in order to pass the criteria, while set pairs with Jaccard coefficient higher than 0.15 can have smaller intersections.

Human verification Process

The final verification of the statistically derived results will be done by human verification, to ensure the validity of the results. This is necessary since we are working with unlabeled text data and would not be able to verify the results without human confirmation. This step verifies the data and enables insights about the performance of the computational tools.

We categorized the text pairs in:

1. University contribution
2. Common content
3. No content match

In the first category, we also included findings about identical topics, which are a University contribution, but to a public entity, or media article or news

about university research, since these are still valid matches for the algorithm, even though the entity might not be the right type.

The human verification was performed on the original text pairs, displaying them as unprocessed texts to simplify the task. [5]

Case And Data Samples

Using the case of the Technical University of Denmark (DTU), we focus on DTU's scientific fields for this investigation. These research areas are dominated by natural and technical sciences due to the university being a technical university, with a focus on applied sciences and Denmark's university with most industry collaborations.

DTU's research was chosen due to data availability and its clear relevance for economic and technical development in its environment. The data availability is driven by fields that have available patents from DTU and/or research collaboration partners.

Previous results suggest that there are more and less relevant scientific fields as units of investigation for knowledge transfer. We determine these with the basic, easier and computationally less expensive methods and focus the more advanced techniques on the more promising/dominant fields. Relevance was defined by the fields expressing a high amount of observed knowledge flow in this and a previous study using text mining and NLP approaches. So for more detailed analyses, where computational performance is crucial, we decided to perform on these advantageous corpora.

Generation of Academic Knowledge Sample

Three distinct text samples represent our sample of university knowledge output. The academic output: scientific publication titles and abstracts including a sample from 2005-2016 [6]. Scientific open access publications including full-texts from 2005-2016[7]. We classify the publications according to their scientific disciplines using the provided department codes in the database. We separate the different research areas (departments), but we have to be aware that the department names and structures changed over time and we have to manually assign them to their scientific fields.

[5] We acknowledge that expert knowledge might in certain cases be required.

[6] Registered and available at the university database ORBIT `http://orbit.dtu.dk/en/`

[7] See footnote 1

Intellectual property rights (IPR) relevant output in our sample patents applied by DTU as the asignee. We keep the usual publications and the open access publications separate since the open access publications have more full-texts and hence the pre-processing and application of the NLP methods will need to be aligned to the available amount of text.

Academic Publications Sample

We have two specific data sets: The first text sample contains all available[8] academic abstracts from the university. Different types/quality of public research knowledge, we have:

1. Academic abstracts of publications:

   - If an abstract of the publication is available
   - If there are more than 10 words in the abstract after the pre-processing
   - If the publication was in one of the 21 relevant departments

   Likewise, we collected all available Open Access publications by university employees from 2005-2016

2. Open Access Publications:

   - Have an available publication full-text that is retrieved by the PURE service
   - Might or might not provide a separate abstract
   - We concatenate the abstracts with the full-texts (might be that we double the abstract text hereby, but this is not always the case and the abstract information is important.

See Table. **??** for the scientific fields and the numbers of publication knowledge available. We have to take into consideration that the amount of open access publications increased highly during the last years. This is relevant, as the knowledge transfer is usually subject to a delay and hence the outcomes between abstracts and open access might not be easy to obtain.

[8] via the registration database

Table 1: Total number of publication abstracts per year (2005-2010 & 2011-2016)

	DP	2005-10	2011-17	Total number
1	Aqua	517	964	1481
2	BioSys	1024	1294	2318
3	ChemBiochem	682	1641	2323
4	Chemestry	686	690	1376
5	CivilEng	737	1272	2009
6	ComputeMath	1470	2411	3881
7	Diverse	1743	874	2617
8	ElectEng	1319	2114	3433
9	EngConSto	116	1087	1203
10	EnviEng	433	1247	1680
11	Food	965	1872	2837
12	MAN	1071	1496	2567
13	MechEng	1173	1818	2991
14	MicroNano	658	1228	1886
15	NUC	5	308	313
16	PhotoEng	1638	2551	4189
17	Physics	551	870	1421
18	Space	474	956	1430
19	Transport	239	621	860
20	VET	540	975	1515
21	Wind	34	1385	1419
	Total abstracts available	16333	28199	44532
	Total entries	40284	37636	77920

The sample containing the most diversity of the actual research outputs from the university is the academic abstract collection. Abstracts contain the most important findings in a very distilled form and are therefore sufficient to trace chunks of knowledge. However, in more complex cases the text material is just not enough to rule out certain overlapping content, which might certainly be related but not necessarily the same research chunk.

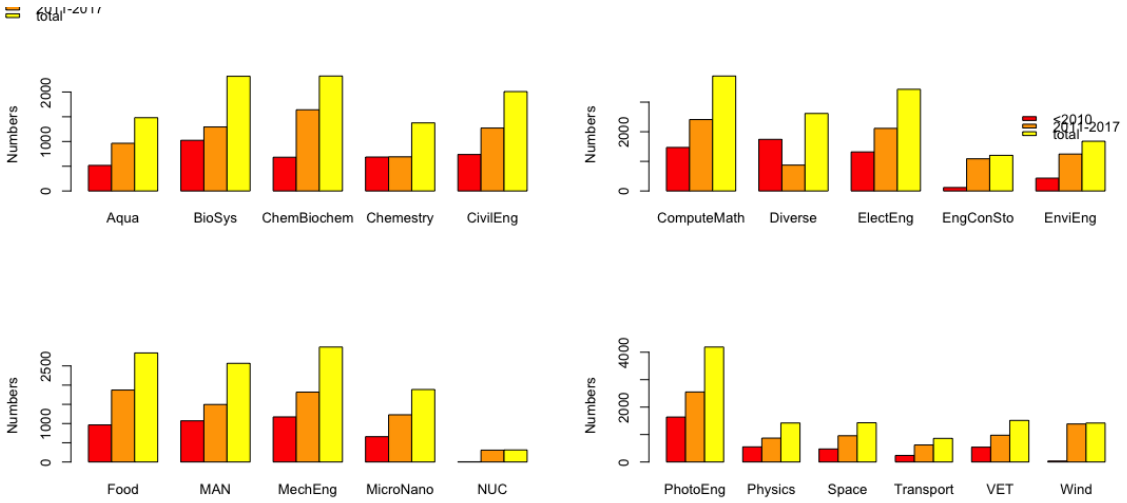


Figure 1: Steps of Abstract department distribution

Table 2: Total Open Access Full-texts available(2005-2010 & 2011-2016)

	DP	2005-10	2011-17	Total number
1	Aqua	189	598	787
2	BioSys	266	628	894
3	ChemBiochem	159	868	1027
4	Chemestry	132	257	389
5	CivilEng	269	726	995
6	ComputeMath	809	1119	1928
7	Diverse	1349	600	1949
8	ElectEng	614	1098	1712
9	EngConSto	22	472	494
10	EnviEng	177	1076	1253
11	Food	206	1442	1648
12	MAN	449	1437	1886
13	MechEng	282	941	1223
14	MicroNano	248	652	900
15	NUC	0	198	198
16	PhotoEng	701	1357	2058
17	Physics	248	431	679
18	Space	125	655	780
19	Transport	115	355	470
20	VET	110	708	818
21	Wind	15	1141	1156
	Total Open Access	6581	17132	23713
	Total entries	40255	37626	77881

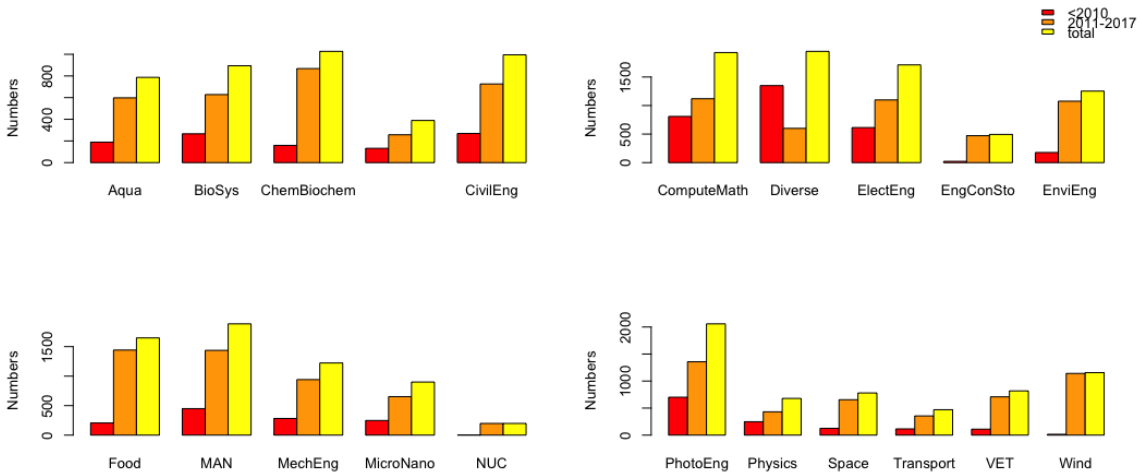


Figure 2: Steps of Abstract department distribution

The number of open access publications increased over the past years and make them a relevant unit of investigation. For our purpose these texts are more explicit than the abstracts, which in some cases, supports the performance of the text mining methods. It especially helps to reduce type I errors (false positives).

Given the sample, we were not able to obtain full coverage of the entire university publications, but the numbers show that a large proportion of the academic publications, produced by the university, have available text data.

University Knowledge in Patents

The true innovative outcome of university research is apparently not only captured in the academic texts they publish, but also in form of IPR protected items like patent declarations. To capture the entire novel knowledge base provided by the university, we add patent from the university to the research knowledge sample. The university patents from 2005-2016 (derived from the ORBIT database. We identified 528 patents assigned. They had to fulfill the following criteria:

- Patent that has DTU as assignee and had to be listed in the DTU database.

- Patent have to be assigned between 2005 and 2016
- It needed an international or European patent number

The publication numbers were retrieved via DTUs own database ORBIT. We decided to include all patents that were either granted by the World Intellectual Property Organization (WIPO) and has a patent publication number with a 'WO' prefix, or by the European Patent Office including patents with a 'EP' prefix in publication numbers. 355 entries had either an assigned or EP patent number or WO, which allowed the retrieval of the full-text patent description and the identification of non-patent literature referenced. The database provided 381 international numbers and 66 European numbers (including instances that have both). We used the following two official databases for the collection of the full-text description:
`https://data.epo.org/expert-services/index-2-3-6.html` the EPO full-text search, but this one provides only full-texts for EP application or publications numbers. To extract the rest, where we only had the WO numbers available we used `https://patentscope.wipo.int/search/en/search.jsf`
For the meta data of individual patents, we used the official Patstat database from 2016.

0.1 Company Websites

This data source provides the foundation for the company knowledge, which we would like to compare to the university knowledge. The texts are gathered from corporate websites of companies that held a contract with the university between 2006 and 2016.

The criteria for companies we include are:

1. Having a national (Danish) company registry number (CVR) or are named *Anpartsselskab* (ApS) (describing limited liability companies in Denmark).
2. Held a contract with the university between 2006 and 2016 or have a hyperlink on the university website [9].
3. Have at least five sub-pages of their website in English.

[9]The list of websites contained many online service platforms, including for example public transportation sites, yellow pages, and firm registries. Large online service providers and social media sites (e.g. Google, Facebook, or YouTube) were excluded from the sample.

This sample was chosen as it constitutes a direct formal link between companies and the university, which is the ideal basis to test and verify the new method [10].

To identify whether a company has a CVR number, we extracted the HTML content of each page along with the CVR if it was available. We fetched the HTML content of the websites using a self designed web-crawler (`https://github.com/nobriot/web_explorer`) and converted the resulting HTML content to plain text and removed any remaining code tags from the text.

Websites were collected between August and November, 2016. We discovered 908,288 total web-pages, meaning single text documents. The total number of companies, which could be identified as collaborators with the university between 2006 and 2016 was 1,225 of which 699 had a CVR number written on their website. Out of these 699 companies, 544 went out of business, underwent mergers, or were renamed[11]. The firms in this sample operate mainly in technology intensive sectors and possess strong R & D divisions.

The number and length of pages varies a great deal between firm websites. Some have an English summary for their main contents, while others (often multinationals) have their entire website in English.

In this HTML text collection, we also ensured to capture PDFs and similar text formats stored on the websites, and converted these formats into raw text to make them usable for analysis. Each website is stored as its own text collection (corpus). These corporal have text size that vary drastically; some contain several thousand web-pages and others just the five minimum pages. Although, though this might seem drastic to store each website as its own corpus, it is an effective way to ensure a comparable and coherent treatment of data and improves the performance of the statistical methods utilized herein.

Results

Pre-processing Relevant Outcomes

The outcomes of the pre-processing revealed some challenges for the academic texts, patent descriptions, as well as website texts. The academic texts (abstracts as well as full-texts) and patents suffered from more or less the same shortcomings, while websites had a different set of issues. Patents and academic texts rely heavily on notations, formulas and other model descriptions

[10] In the future, however, we plan to expand the sample.

[11] We tried to identify the new names or entities, however this was not possible in all cases.

that often contain numbers and special characters, which are removed in the pre-processing. This is crucial for co-word occurrence and limits the findings, even though many of the instances actually are generated into the same character string. In particular, comparing patents with academic texts has shown that the pre-processing reduction results in coherent strings in both texts in most cases. The websites on the other hand rely less heavily on scientific notations, but contain other challenges. Some information is not given in text form and is embedded in texts as images or videos. Currently these are only detected as code tags in the HTML texts and fully lost in the process. In some rare cases broken code tags remain in the cleaned texts and prevent the identical deconstruction similar to the one of the university texts. In this case we did not yet find a way to identify the matching strings. However, websites have a comparatively high usage of proper nouns, product names and similar terms, that help to identify common methods, research and inventions, hence some terms seem like the result of unsuccessful pre-processing, but are, in reality, just a representation of names, which are shrunk to a string of characters. Overall the manual inspection of the pre-processed texts shows the difference between the retrieved documents.

Text To Text Comparison

Table 3 shows a summary of TFIDF-TFIDF comparison results. The total number of matches are indicated in brackets while the other number specifies the number of confirmed matches after human validation.

Table 3: Comparisons made TFIDF keywords

Comparison TFIDF	Comparison Nr.	Matches Nr. (pre clean)	Jaccard Max.	Jaccard Threshold
Patents vs Abst.	15.808.860	46 (46)	0.27	0.2
Patents vs OpenA	8.250.911	44 (53)	0.45	0.25
Patents vs Web	49.185.961	23 (56)	0.14	0.1
DTU vs Web	6.137.022.288	91 (124)	0.24	0.13

LDA & TFIDF

Table 4 shows a summary of TFIDF-LDA comparison results. The total number of matches are indicated in brackets while the other number specifies the number of confirmed matches after human validation.

Table 4: Comparisons made TFIDF & LDA keywords

Comparison TFIDF vs LDA	Comparison Nr.	Matches Nr. (pre clean)	Jaccard Max.	Jaccard Threshold
OpenA vs Patents	1.162.101	31 (33)	0.23	0.15
Abst vs Patents	2.109.351	48 (48)	0.21	0.15
Web vs Patents	6.927.601	13 (19)	0.14	0.1

Results Per Corpora

Results details are given in the following subsections for the comparisons between the different text type pairs.

Patents & Academic Abstracts

In the first step, we compared the retrieved keywords from the TFIDF indexing from the university patent descriptions (355) and the university abstracts (44.532). In the second step we used additionally the keywords of the LDA topics from the patent descriptions to the TFIDF keywords from the abstract corpus.

The first step is useful to confirm the relatedness between the university publications and patents and at the same time to assess the performance of this simple algebraic method. The TFIDF method is applied directly on a text to text basis. Additionally, it only extracts observable features, unlike the LDA, and one can therefore assume that if the features are found in both texts they will be found to be relevant.

However, in our case it is pertinent to determine whether abstracts tend to have enough information to allow just the use of abstracts or a full-text analysis is inevitable, which is computationally much more expensive. The comparisons gave 46 matches and 30 of them were found to be clearly related papers. This means that we could only identify a fragment of related literature for the patents, but the found matches were comparatively well identified and found papers that are based on patented inventions. Common terms show a good quality in overlapping terms, which are rather specific and content relevant.

In the second step, we used an LDA with $K = 50$ to derive overall topics[12].

[12]It is a comparatively high number of topics for a small corpus of 355 documents, but we follow the assumption of the novelty of patents and therefore believe that the diversity in topic and contents is much higher than with other forms of texts

Table 5: Example common terms text pairs (Patent vs Abstracts)

	Common terms				
Example match one	ceram densif	presint insitu	cast porous	barrier sofc	cgo electrolyt
Example match two	mwm chamber	microorgan mfcs	submers cell	electrod cathod	voltag fuel

The 48 matches were obtained from the comparison between the patent topics and the TFIDF from the abstracts. The low number of keywords for the abstracts (compared for instance to full-text publications) resulted here in comparatively low Jaccard Similarity scores and at the same time in a low common term count. As in the previous comparison no concrete match could be identified, but the common topics were found narrowing future analytically steps down.

Table 6: Common terms for one text pair patent LDA and abstracts

Common terms				
surfac	catalyt	area	support	show
activ	method	high	temperatur	calcin
catalyst	acid	prepar	ammonia	scr

Patents & Academic Open Access Publications

Comparing the open access publications from the university (23713 full-texts) with the patent descriptions (355) we, again, used two methods; first the TFIDF keyword comparison and second the LDA topic keyword comparison from the patents to the TFIDF keywords of the open access publications.

44 relevant hits were found with the TFIDF method only. This is a low number considering the 355 patents, and we assume that most of the inventors are also researchers that publish papers related to or about their invention. Out of these 44 matches, we found 40 to true matches containing the same invention, 4 category 2 (identical topic but different invention) and none category 3 (not related at all). So the type II error is extremely low and hence shows that the performance on full-texts, in this particular combination, is excellent.

When comparing with the findings of the Patents vs Abstracts, we found 7 out of these 44 matches to be the identical or common matches, which link the same publication number and patent number. These 7 common matches

were in both cases (Patent vs Abstracts and Patent vs full-texts) classified as rank 1 matches.

4 of the matches actually refer to the same publication number, but associated with a different patent. Leaving only 33 unique new full-text publications matches, of which 29 were also in the abstracts data and not identified as matches. This suggests a rather high type I error rate for the abstracts in addition to the already high type II error rate.

In the case of the full-text publications, the 12 matches (rated as category 1) were identified via the abstract and of these 7 were also detected by the full-texts. This suggests a better performance of the patent vs full-text instance.

Table 7: Common terms for one text pair

Common terms				
sialidas	sialyllactos	mutant	motif	
longum	cgmpbound	acid	amino	mutat
tcts	imo	perfringen	oligosaccharid	vtnkkkq
gos	transsialyl	trsa	glycan	hydrolas
sialic	transsialidas	enzym	acceptor	cgmp
infanti	sialyl	lactulos	cruzi	prebiot
trypanosoma	mugal			

In comparison to the Patent vs Abstracts, we observe here longer common keyword lists, which make the hits more precise and allow for higher Jaccard similarity (up to 0.44 Jaccard similarity) and extensive term co-occurrence.

Secondly, we combined the LDA of the patent with the TFIDF indexing of the full-text publications, following the general idea that the topics in a rather small but diverse corpus, like the one of the the patents, might derive the underlying structures in a way that allows us to match them with the single keyword lists derived for the academic corpora. We found 31 relevant hits, since the LDA does not translate directly into a single document, as for instance the TFIDF application, it was needed to retrieve the paper for the topic that have the highest probability. So for instance, if the topic number 10 has a match with a full-text publication, we retrieved the 2 most relevant (or probable) documents for a given topic number[13].

The general overlap shows clear topic overlap in the corpora and is therefore

[13]Another possibility would have been to derive the documents that represent the common terms that actually made the match, however since the keyword and topic word lists were not extremely extensive this did not seem necessary.

promising for further statistical analysis. It is helpful to identify fields with overlapping knowledge or notions.

Table 8: Common terms for one text pair

Common terms				
loop	semiconductor	convert	feedback	coupl
embodi	signal	rectifi	piezoelectr	transform
power	reson	switch	capacitor	circuit
input	electrod	output	voltag	primari

Patents & Websites

The comparison between websites and the academic patents required generally great adjustments of the Jaccard similarity threshold, as the co-word occurrence is really low. The retrieved term matches show common fields, but are too little to extract actual knowledge chunks that are related. Out of the 23 matches only one clear hit was identified, which is a very low performance for the method, However, with a more extensive sample of patents and companies this might still be an approach to consider.

Table 9: Common terms list for two text pairs (patents vs websites)

Common terms					
Example match one	substrat	genomescal	metabol	aerob	biomass
	flux	stoichiometr	silico	atp	glucos
Example match two	infect	aeruginosa	biofilm	antibiot	clinic
	microbi	treat	ulcer	bacteri	cystic

The poor matching rate between patent and websites suggests that patents details might now always be present on websites using a patent knowledge or invention. Commercial application of a patent may results in many different end products and it would not be expected to be able to find the relation between the the website and the patent in this case.

University Knowledge & websites

The comparison between university knowledge and corporate websites is in many aspects different from the previous comparisons, due to linguistic composition and the purpose the texts serve. For computational efficiency, we took both the abstracts of the patents and the publication abstracts and compared

this corpus to the websites. We used for this instance the TFIDF keyword comparison. The comparison identified around 23 matches, which were rated as category 1, while the vast majority (46) of the given matches were only topic related and another quarter did not show any content relatedness. However, for this specific example we derived a comparatively high number of matches (91 after cleaning up) setting the Jaccard Similarity threshold comparatively low to not miss any potential matches. All, but one, of the rated category 1 text pair matches were in the higher end of the Jaccard similarity and hence the performance reducing the type I error could have been significantly improved by setting a higher threshold and derive only the first 40 hits.

Additionally, it has to be taken into consideration, that the expectations of overlap have to be adjusted and must be lower than the overlap between research (patents and publications) of the same university. The difference in Jaccard scores and threshold can be seen in Table 3. The hits found in this particular case were rather unrelated and suggest that the sample might be too small for a valid verification of this particular method.

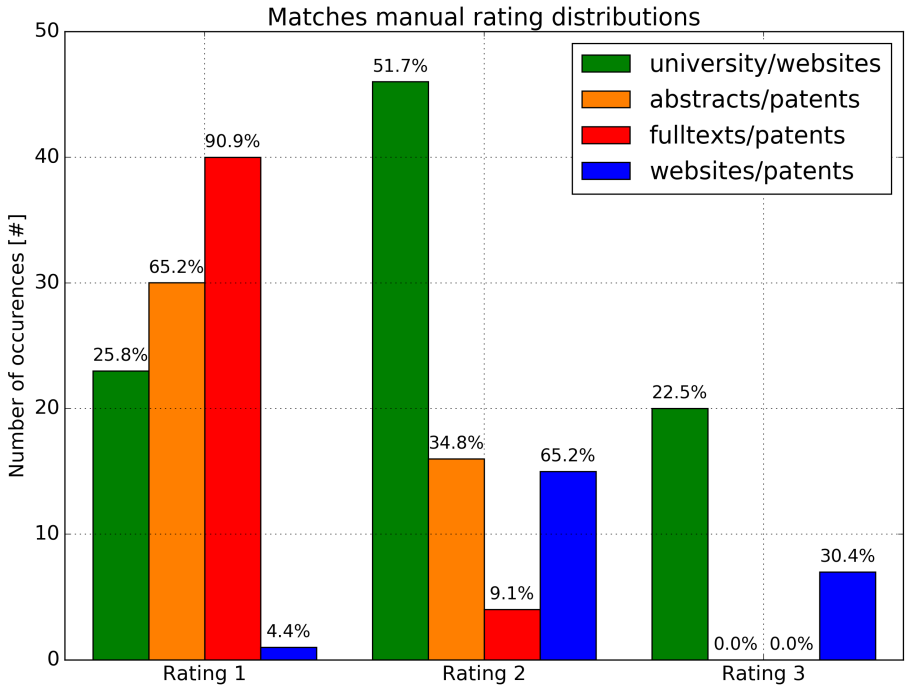


Figure 3: Match rating distributions for the different methods

Findings

This final section is divided into different aspects of our main findings, including technical as well as conceptual insights, potential technical improvements, and the outlook for future research in this particular area.

Technical Insights

The statistical comparison between the patents, abstracts, and open access full-texts suggest that the TFIDF comparison works well comparing linguistically similar text types. However, comparing the results from the comparison of patent TFIDF keywords against abstracts to full-texts clearly shows that in this case abstracts are not a valid alternative. This is highly relevant for contemporary research in several fields, that make use of bibliometrics or text mining, since most of them rely solely on titles or abstracts (Bornmann et al. 2017, Zhang et al. 2010). The use of titles and abstracts may in some cases be appropriate, but might, as shown in this study, not be as robust as using the full-text analysis, which are often not as easy to compile or analyze. The detection of transfer using websites is promising. However, statistical adaptation was necessary to identify the matches.

The results suggest that the TFIDF indexing method produces less than optimal results and additional adaptation is necessary to improve the outcomes for this specific comparison of websites

Given the combined results of the LDA and the TFIDF comparisons, we can clearly state that the keywords related to most matches show clear tendencies towards particular research areas, or patent fields. This marks clearly the scientific areas in which the common ground are most dominant.

Conclusion

Our findings suggest that academic texts, in particular full-text, and patent descriptions can very well be used to trace knowledge chunks. For instance, it is possible to identify the concrete papers related to a specific patent, as well as to trace this knowledge further in less related texts like websites. However, given the scope and limitations of this paper, how far these traces in comparison go will have to be assessed via a larger scale study.

Surprisingly,our results suggest that it is not much more likely that patent content gets transferred than usual publications. This is crucial, since in many studies the existence of a patent is used synonymously with knowledge/technology transfer and seen as the key indicator for university contri-

bution. This might in deed still hold up, but publications might be as good as an indicator, at least in some scientific fields.

For the future, it might be suitable to identify the related publications of a patent and use their references as alternative to the non-patent citations in the patents. They might provide much more insight, since they are added according to scientific necessity and not for legal reasons as is often done for non-patent literature. Furthermore, these references could serve as a basis for more extensive knowledge corpora, which would increase the chances of detecting knowledge chunks.

Limitations

Our findings are in the early stage and need more testing with larger and more diverse data sets. In particular the conclusions about patented inventions transfer to the industry need to be verified with larger and more diverse data sets. Unfortunately, the ability to assess the actual number of type II errors is limited, and this likely will continue to be an issue in future work. The human verification process is a clear limitation of the method and will have to be at least reduced in future iterations.

References

Agrawal, A. & Henderson, R. (2002), 'Putting Patents in Context: Exploring Knowledge Transfer from MIT', *Mgmt. Sci.* **48**(1), 44–60.

Agrawal, A. K. (2001), 'University-to-industry knowledge transfer: literature review and unanswered questions', *International Journal of Management Reviews* **3**(4), 285–302.

Arundel, A. & Bordoy, C. (2008), 'Developing internationally comparable indicators for the commercialization of publicly-funded research'.

Berry, M. W. & Castellanos, M. (2007), 'Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition', p. 241.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Bornmann, L., Haunschild, R. & Hug, S. E. (2017), 'Visualizing the context of citations referencing papers published by eugene garfield: A new type of keyword co-occurrence analysis', *arXiv preprint arXiv:1708.03889* .

Branstetter, L. (2006), 'Is foreign direct investment a channel of knowledge spillovers? Evidence from Japan's FDI in the United States', *Journal of International Economics* **68**(2), 325–344.

Chapman & Hall/CRC (2010), *Handbook of Natural Language Processing, Second Edition.*

Cheah, S. (2016), 'Framework for measuring research and innovation impact', *Innovation* **18**(2), 212–232.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011), 'Natural Language Processing (almost) from Scratch', *The Journal of Machine Learning* **12**, 2493–2537.

D'Este, P. & Patel, P. (2007), 'University-industry linkages in the UK: What are the factors underlying the variety of interactions with industry?', *Research Policy* **36**(9), 1295–1313.

Drucker, J. & Goldstein, H. (2007), 'Assessing the Regional Economic Development Impacts of Universities: A Review of Current Approaches', *International Regional Science Review* **30**(1), 20–46.

Franceschini, S., Faria, L. G. D. & Jurowetzki, R. (2016), 'Unveiling scientific communities about sustainability and innovation. A bibliometric journey around sustainable terms', *Journal of Cleaner Production* **127**, 72–83.

Griffiths, T. L. & Steyvers, M. (2004), 'Finding scientific topics.', *Proceedings of the National Academy of Sciences of the United States of America* **101 Suppl**, 5228–35.

Grimpe, C. & Hussinger, K. (2013), 'Formal and Informal Knowledge and Technology Transfer from Academia to Industry: Complementarity Effects and Innovation Performance', *Industry and Innovation* **20**(January 2016), 683–700.

Grün, B. & Hornik, K. (2011), 'topicmodels : An R Package for Fitting Topic Models', *Journal of Statistical Software* **40**(13), 1–30.

Gulbrandsen, M. & Slipersaeter, S. (2007), The third mission and the entrepreneurial university model, *in* 'Universities and Strategic Knowledge Creation: Specialization and Performance in Europe', chapter 4, pp. 112–143.

Heinze, N. & Hu, Q. (2006), 'The evolution of corporate web presence: A longitudinal study of large American companies', *International Journal of Information Management* **26**(4), 313–325.

Howells, J. (2002), 'Tacit Knowledge, Innovation and Economic Geography', *Urban Studies* **39**(5-6), 871–884.

Huggins, R., Johnstona, A. & Steffensonb, R. (2008), 'Universities, knowledge networks and regional policy', *Cambridge Journal of Regions, Economy and Society* **1**(2), 321–340.

Lissoni, F. (2012), 'Academic patenting in Europe: An overview of recent research and new perspectives', *World Patent Information* **34**(3), 197–205.

Liyanage, C., Ballal, T., Elhag, T. & Li, Q. (2009), 'Knowledge communication and translation - a knowledge transfer model', *Journal of Knowledge Management* **13**(3), 118–131.

Lundberg, J., Tomson, G., Lundkvist, I., Skår, J. & Brommels, M. (2006), 'Collaboration uncovered: Exploring the adequacy of measuring university-industry collaboration through co-authorship and funding', *Scientometrics* **69**(3), 575–589.

Mao, W. & Chu, W. W. (2007), 'The phrase-based vector space model for automatic retrieval of free-text medical documents', *Data and Knowledge Engineering* **61**(1), 76–92.

Paukkeri, M.-s. & Honkela, T. (2010), 'Likey : Unsupervised Language-independent Keyphrase Extraction', (July), 162–165.

Polanyi, K. (1962), 'Personal knowledge: Towards a post-critical philosophy", chigago university press, chigago'.

Ponweiser, M. (2012), Latent Dirichlet Allocation in R, PhD thesis.

Roach, M. & Cohen, W. M. (2013), 'Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research', *Management Science* **59**(2), 504–525.

Stahl, M. J., Leap, T. L. & Wei, Z. Z. (1988), 'Publication in Leading Management Journals As a Measure of Institutional Research Productivity', *Academy of Management Journal* **31**(3), 707–720.

Sung, T. K. & Gibson, D. V. (2000), 'Knowledge and Technology Transfer : Levels and Key Factors', *Proceeding of the 4th International Conference on Technology Policy and Innovation* .

Toutkoushian, R. K., Porter, S. R., Danielson, C. & Hollis, P. R. (2003), 'Using publications counts to measure an institution's research productivity', *Research in Higher Education* **44**(2), 121–148.

Vincett, P. S. (2010), 'The economic impacts of academic spin-off companies, and their implications for public policy', *Research Policy* **39**(6), 736–747.

Walsh, J. P., Lee, Y.-N. & Jung, T. (2016), 'Win, lose or draw? the fate of patented inventions', *Research Policy* **45**(7), 1362–1373.

Xia, T. & Chai, Y. (2011), 'An improvement to TF-IDF: Term distribution based term weight algorithm', *Journal of Software* **6**(3), 413–420.

Zawdie, G. (2010), 'Knowledge exchange and the third mission of universities', *Industry & Higher Education* **24**(3), 151–155.

Zhang, G., Xie, S. & Ho, Y.-S. (2010), 'A bibliometric analysis of world volatile organic compounds research trends', *Scientometrics* **83**(2), 477–492.

Zhang, Y., Zhang, G., Chen, H., Porter, A., Zhu, D. & Lu, J. (2016), 'Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research', *Technological Forecasting and Social Change* **105**, 179–191.