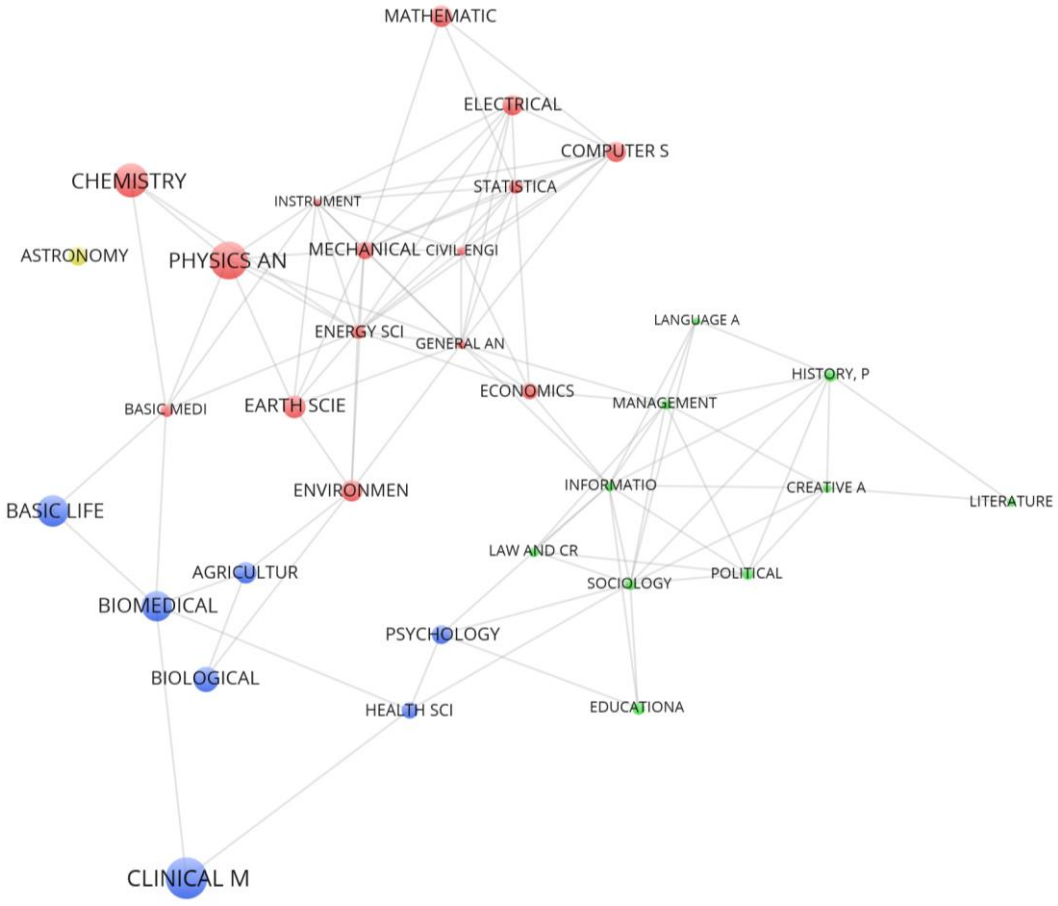
Content words as measure of structure in the science space

Towards a classification of publications based on noun phrase occurrence



A master's thesis by Wout S. Lamers BSc

Content words as measure of structure in the science space

Towards a classification of publications based on noun phrase occurrence

Submitted as part of the master's programme innovation sciences at Utrecht University

Author:	Wout S. Lamers
Student id:	3348377
Email:	w.s.lamers@students.uu.nl / wout.lamers@gmail.com

Thesis committee:

University supervisor:	Dr Gaston Heimeriks	(Utrecht University)
External supervisor:	Dr Ingeborg Meijer	(Centre for Science and Technology Studies)
External supervisor:	Dr Ed Noyons	(Centre for Science and Technology Studies)
Second reader:	Dr Jarno Hoekman	(Utrecht University)

Project duration:
31-10-2014 – 18-09-2015

Completed at:

Universiteit Utrecht

Utrecht University
Heidelberglaan 8
3584 CS Utrecht
The Netherlands

Centre for Science and Technology Studies
Wassenaarseweg 62A
2333 AL Leiden
The Netherlands

Content words as measure of structure in the science space

Towards a classification of publications based on noun phrase occurrence

Abstract

Recent developments in the production of scientific knowledge, namely an increase in interdisciplinary scientific research and interaction between academia and industry, pose challenges to citation-based methods for studying the structure of the science space. As an alternative, we ask the question whether it is possible to construct maps of science and find disciplinary similarity structures based not on citation, but on the content of publications' titles and abstract. We present a theoretical framework in which we define disciplines as being distinct from one another based on their associated cognitive elements. Specifically, each discipline will have its own unique vocabulary used when communicating its research results in the form of publications. Linking this framework to text processing methods, we elaborate on how the occurrence of noun phrases within disciplines may be used to represent disciplines and documents as term-occurrence vectors in a high-dimensional vector space model.

From the Web of Science database, we collect over seven million publications spread over 33 disciplines. Comparing the angles between these disciplines' term-occurrence vectors we construct a discipline similarity structure and use this structure to generate maps and a clustering solutions. We find that both the structure and the clusters are highly stable over time. We explore two different ways of computing a relevance score for noun phrases. One may be useful at finding discipline-specific cognitive content, the second is highly effective at removing low-relevance noun phrases from the vector space model while preserving the similarity structure. The effects of this pruning of low-relevance terms is further explored in the final experimental step of the research, where we divide the sample in a test and training sample and classify 1.4 million test publications based on their highest similarity to the training sample disciplines. We find encouraging classification performance, nearly similar with and without pruning of low-relevance terms.

Our results indicate that we can indeed derive a stable and meaningful structure of the science space from publications' title and abstract text. The classification shows that this structure can subsequently be put to use to place new publications into this structure with encouraging accuracy. This is an important conclusion, as so far methods for mapping the science space have been mostly restricted to citation data. These results open new avenues for research, potentially into the systematic assessment of novelty and new combinations in science.

Preface

This thesis is the product of an almost yearlong research project at Utrecht University and the Leiden Centre for Science and Technology studies, during which it evolved from the slightest hint of an idea into what I feel is a fitting conclusion to my time as a student. Writing this thesis has been a challenge, frustrating at times, but overall incredibly rewarding and a great experience. It is a product born out of genuine interest in the structure and dynamics of the world we live in, and while the *structure of the science space* may sound like an obscure research topic at best to many, I believe it is a fascinating area of research that I enjoyed contributing to.

I would like to take this opportunity to express my sincere gratitude to the people who made this thesis possible. First and foremost, my supervisors, Dr Gaston Heimeriks, Dr Ingeborg Meijer and Dr Ed Noyons, who gave me the opportunity to conduct my research and write this thesis at a top-of-the-line research institute. They offered me both the freedom to follow my interests and shape this research project the way I envisioned it, as well as the guidance and feedback that helped bring focus and refinement to my research. The second reader, Dr Jarno Hoekman, whose astute comments on my research proposal were invaluable in shaping the research into a coherent whole. My colleagues at CWTS, who welcomed me so warmly, made this thesis a pleasure to work on, and allowed me to experience what scientific research is really like. The other teachers who have imparted their knowledge on me over the years. And finally, my friends and family, for their support, encouragement, for providing distraction when I needed it, and for all the experiences that we shared over the years.

Contents

1: Introduction

Over the past few decades, several authors have argued that profound changes are taking place in the way new scientific knowledge is produced. Publications such as *The New Production of Knowledge* (Gibbons et al., 1994) and subsequent alternatives such as the *triple helix* model of industry, academia and government interaction (Leydesdorff & Meyer, 2006) have set off a debate on whether the disciplinary divisions in the science system are still an accurate representation of reality. Increasing collaboration between science and industry to address complex socio-scientific issues is prompting changes in the ways researchers and policymakers alike think about performance measures and evaluation of scientific output (Wagner et al., 2011). Subsequently, interdisciplinarity now plays a role both in the allocation of research funds as well as in the assessment of social impact of research. Research crossing the borders of traditional scientific disciplines or transcending the divide between science and industry is seen as a potent source of new combinations from formerly separate knowledge pools.

From an innovation sciences perspective, the increase in interdisciplinary scientific research can be regarded in a Schumpeterian sense. According to Schumpeterian theory, innovation is the result of *neue Kombinationen* – new combinations – of existing knowledge (Kurz, 2012; Schumpeter, 1934). Combining knowledge from different scientific disciplines, as well as industry, should pave the way for new combinations previously not possible within isolated disciplines. Indeed, new emerging research fields, for instance nanotechnology and bioinformatics, could be considered the result of combinations of established disciplinary knowledge. Thus, scholars of innovation might argue that increasing inter- and transdisciplinary collaboration should have a positive effect on the production of new scientific insight.

From a scientometrician's perspective though, these developments pose significant challenges. Interdisciplinarity is a hotly contested topic, with a myriad of associated concepts and even more interpretations of these concepts. In 1996, Hicks and Katz noted that the lack of measurements with regard to interdisciplinarity is "not surprising given the apparent impossibility of even agreeing on a definition of interdisciplinary research" (Hicks & Katz, 1996, p. 387). Debate on a proper scholarly working definition of the relevant concepts continues to this day (e.g. Porter, Roessner, Cohen, & Perrault, 2006; Schmidt, 2008, 2011). The take-away message is that increasing collaboration, both between and beyond scientific disciplines, increases the complexity of the science system.

The traditional tools used for gaining insight into the science system and its structure are various types of bibliometric mapping techniques, constructed most frequently using various types of citation data. For instance, *global maps of science* have been created at the journal level based on the Web of Science subject categories (Leydesdorff, Carley, & Rafols, 2013; Leydesdorff & Rafols, 2009) as well as at the publication level using direct citation relations (Waltman & van Eck, 2012). The maps produced by both methods allow researchers to decompose the body of science into disciplinary and subdisciplinary structures. However, given that they rely on citation data for the positioning and clustering of nodes, they are limited in the sense that these disciplinary structures cannot subsequently be applied to classify, for instance, research proposals and grey literature, or any type of texts that do not (yet) have complete citation records or adhere to academic citation standards. Alternative and secondary types of research output are becoming more and more relevant as scientists focus on addressing socio-scientific issues and collaborating with industry. This is further illustrated by a recent rise of alternative measures for scientific impact that move beyond pure citation based metrics. For instance, recent developments include *social media metrics* or the broader *altmetrics*, for article level metrics or 'alternative' metrics (for a detailed discussion on altmetrics see Costas, Zahedi, & Wouters, 2014).

The increasing interaction between science and non-scholarly research and the increasing complexity of the science system raises concerns about the extents to which citation data can continue describe the science system as it evolves further. On top of this, an exclusive reliance on citation data ignores a trove of other information that one could use to find structure in the scientific landscape. Methods combining co-occurrence of title words and references seem promising for the purpose of generating maps of publications within limited journal sets representing research fields (e.g. van den Besselaar & Heimeriks, 2006) and description of research topics of citation clusters (e.g. Braam, Moed, & van Raan, 1991), but have not yet been employed to map the overall structure of science to the extent that citation-only methods have. Furthermore, their continued reliance on citation data makes them an interesting extension to citation-data-only mapping methods, but not a replacement capable of finding structure when such citation data is lacking.

This brings us to the primary goal of our research. So far, methods for generating maps of science and for finding structure in the science space hinge on citation data, which is simply not present in many new types of research output. To create a more complete picture of the complexity of the science system, eventually secondary research outputs will have to be incorporated somehow. We turn our hopes to language itself, which remains the primary way of communicating research results. Content words have already been used to contribute to mapping the science space – but rather than a supporting role, can they be used as the primary data source from which we derive structure in the science space? The aim of this thesis, then, is to develop and explore fundamentally this alternative method for finding structure in science and determining publications' place within this structure, relying not on citation data but on the occurrence of content words in publications' title and abstract. The starting point for this research is the Web of Science database and a set of pre-defined disciplines. The central question posed in this thesis is as follows:

Can we find a consistent and meaningful structure in science, and documents' place within this structure, using publications' title and abstract text?

In order to answer this question, we will discuss the relevant background to this question in the following chapter, wherein we review current notions of disciplinarity and interdisciplinarity as well as existing methods for mapping the science space. In the third chapter we derive from this background an abstract theoretical framework for disciplinarity and interdisciplinarity, which allows us to understand the relationships between some of the most important concepts used when discussing these notions. This chapter concludes with expanded research subquestions derived with the help of our theoretical framework. These subquestions and the theoretical framework serve as the foundation on which we develop our method for comparing disciplines and documents, detailed in the fourth chapter. At its heart, the method relies on identifying the occurrence patterns of cognitive elements – more specifically noun phrases, sequences of words consisting exclusively of nouns and adjectives ending in nouns – at the discipline level. A distance metric can be computed for each pair of disciplines based on the similarity of their noun phrase occurrence, and likewise publications' similarity to disciplines may be established based on the noun phrases occurring in their title and abstract.

The fifth chapter begins the experimental portion of the thesis and contains the result of our exploratory investigation of the sample. In its first subsection we discuss how we used over seven million publications extracted from the Thomson Reuters Web of Science database in the period 2000-2010 to generate a map of disciplines, based on the similarity of their overall noun phrase occurrence patterns. In the second subsection we segment the sample on a per-year basis and investigate whether the landscape formed by our discipline similarities changes over time. The third subsection delves deeper still, visualizing and describing changing noun phrase occurrence patterns within disciplines over the years. In the final subsection of the fifth chapter, closing our analysis of the overall structure

uncovered by the method, we adapt the method to use cited references instead of noun phrases, and compare the resulting disciplinary structure with the one found using noun phrase data, to see how the noun phrase method compares to more traditional citation-based methods.

Up until there, our focus has been on the overall occurrence patterns of noun phrases within disciplines. In the sixth chapter, we discuss ways to determine exactly which noun phrases are most relevant to each discipline, and compute disciplinary relevance scores for each noun phrase. We further present and discuss the most relevant noun phrases for a select amount of disciplines. Finally, in the seventh chapter, we divide the data into a training sample and a test sample and use the former to classify the latter, in order to validate the method and verify that it can be used to find publications' disciplinary association. Two means of classifying the test sample are compared: one using all the available noun phrase data in the training sample, another using only those noun phrases with high relevance scores in at least one training sample discipline. This thesis closes with chapters containing our final conclusions, and a discussion and reflection as well as recommendations for further research.

The contribution this research aims to make is highly relevant to the sciences. Maps of science provide valuable insight into the relative positions of research activities and help researchers, research institutes and funding agencies to position themselves in this structure. We go beyond this, aiming to develop a method for directly relating text to scientific research fields based on their noun phrase usage. If successful, this contributes to making methods of mapping the science space more futureproof and capable of handling novel forms of research dissemination which may rely less on traditional citation. Furthermore, being able to find the cognitive roots of new pieces of knowledge, and knowing how those cognitive roots or topics relate to one another, will allow for a more accurate assessment of the novelty value of scientific contributions. This is especially important because while interdisciplinary scientific research has been heralded as the way to address complex socio-scientific problems, measuring and evaluating the true impact of interdisciplinary scientific research has proven difficult. This, in essence, also makes our research relevant to society, for if interdisciplinary research is to solve the larger problems facing humanity today, we all benefit from a more accurate assessment of that research.

2: Background

In this chapter we discuss the relevant theoretical background of the notions of disciplinary and interdisciplinary science, as well as prominent means of mapping the science space, as basis for constructing our own theoretical framework in the next chapter.

2.1: What is a discipline?

If we hope to map the disciplinary structure of science, we must first establish what we mean when we use the word "discipline". While there is no doubt that disciplinary structures exist in science (see, for instance, the organisation of universities into faculties and departments, and the prevalence of journals devoted to limited sets of topics), there is no single agreed-upon definition of what constitutes a discipline or its boundaries (Wagner et al., 2011).

Some authors define disciplines as bodies of knowledge (e.g. Alvargonzález, 2011), as domains characterised by a distinct central problem (Darden & Maull, 1977; Porter et al., 2006), or as scientific communities (e.g. Lélé & Norgaard, 2005). In a chapter in *Practicing Interdisciplinarity* (Stehr & Weingart, 2000), Stephen Turner defines disciplines specifically as groups of degree-holders and degree-granting units (Turner, 2000), while in the same book Peter Weingart states that a discipline is a diffuse social organisation for the production of knowledge (Weingart, 2000). Van den Besselaar and Heimeriks (2001) state that a disciplinary research field is "*a group of researchers working on a specific set of research questions, using the same set of methods and a shared approach*" (van den Besselaar & Heimeriks, 2001, p. 2). Similarly, Wagner et al. (2011), following Porter et al. (2006), define a discipline as "*having a central problem with items considered to be facts relevant to that problem, and having explanations, goals, and theories related to the problem*" (Wagner et al., 2011, p. 15). As noted by Van den Besselaar & Heimeriks (2001), the concept discipline seems related to Thomas Kuhn's scientific paradigms in the sense that disciplinary research is normal problem solving within a paradigm (see Kuhn, 1970).

Attempting to finalize the debate on the definition of *discipline* is beyond the scope of this thesis. Instead, I will use the above definitions to arrive at a broad definition of the concept, which is adequate for the purpose of this research. The various definitions of what constitutes a discipline outlined in the previous paragraph involve both cognitive and social elements. As observed by Ed Rinia (2007), disciplines comprise codified knowledge, agreed-upon methods, and a common language, all of which can be taught to those who wish to enter the discipline. Disciplinary research is practiced by a community of researchers in the context of organisational structures, such as university faculties and departments, journals, and reward mechanisms. I follow Rinia (2007) in defining disciplines as structural features within the larger system of science, distinct from one another by their cognitive and social dimensions. Cognitive dimensions include such things as objects of interest, accepted knowledge, agreed-upon methods and a common terminology shared by a community of academic peers, while social dimensions are the features along which such a peer community is organized, such as university faculties and departments, conferences, journals, but also reward and reputation structures and validation functions. Consequently *disciplinary scientific research* (or monodisciplinary or unidisciplinary research) is research conducted within the cognitive and social boundaries of a discipline. This definition is broad enough not to be at odds with the various definitions used by other scholars described in the previous paragraph, but specific enough to allow for a delineation between disciplines based on their cognitive or social features – and while boundaries may be difficult to observe directly, they can be established indirectly by identifying the disciplinary cognitive or social elements contained within them.

2.2: Towards a continuum of disciplinary integration

A discussion of disciplinarity in science would not be complete without considering research that does not fit within the boundaries of a single discipline. If disciplinary research is normal problem solving, then non-disciplinary research must be exceptional somehow. Indeed, Gibbons et al. (1994) contrast traditional *mode 1* science, which is disciplinary and focused on academic knowledge production, to emerging *mode 2* science, which is transdisciplinary in the sense that its focus lies on larger socio-scientific problems, necessitating the mobilization and integration of theories and methods from different fields. The notion that the complexity and scope of problems is a driving force behind non-disciplinary research is found in multiple publications (e.g. Hicks & Katz, 1996; Porter et al., 2006; Schmidt, 2008, 2011; van den Besselaar & Heimeriks, 2001; Wagner et al., 2011), but at the same time claims of research transcending disciplinary boundaries and addressing complex societal issues may very well be the result of scientists' drive to legitimize their work (Weingart, 2000). Furthermore, multiple scholars describe the practice of "borrowing" where researchers working in disciplinary context adopt a method from a second discipline without also committing to that second discipline's objects of inquiry (e.g. Porter et al., 2006). This limited crossover between disciplines may lead one to suspect that the distinction between disciplinary and non-disciplinary research is not quite as black and white as it is sometimes made out to be. Indeed, according to Hessels & Van Lente (2008), *mode 1* and *mode 2* are best regarded as ideal types or extremes on a continuum of scholarship rather than exclusive and definitive modes of knowledge production.

If *mode 1* monodisciplinary and *mode* 2 transdisciplinary research are two extremes on a scholarly continuum, there must exist intermediary forms of non-disciplinary research. Once again, multiple different concepts appear in literature, but this time definitions vary only subtly among authors, at least as far as the most prevalent concepts – *multidisciplinary* and *interdisciplinary* research, as well as the previously mentioned *transdisciplinary* research – are concerned.

Multidisciplinary research entails approaching a subject from multiple disciplinary angles (van den Besselaar & Heimeriks, 2001), with researchers working independently or sequentially on a common problem (Choi & Pak, 2006; Rosenfield, 1992; Stokols et al., 2003), and at most drawing on knowledge or methodology from different disciplines but not attempting to unify them (Alvargonzález, 2011; Choi & Pak, 2006).

Interdisciplinary research "analyses, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole" (Choi & Pak, 2006, p. 359). It combines disciplinary approaches into its own methodology (van den Besselaar & Heimeriks, 2001) wherein researchers work jointly together but still from their own disciplinary perspectives on a basis of partnership (Rosenfield, 1992; Stokols et al., 2003; Wagner et al., 2011).

Transdisciplinary research, finally, is based on a shared conceptual framework (Rosenfield, 1992) and mutual interpretation of disciplinary epistemologies (Gibbons et al., 1994; van den Besselaar & Heimeriks, 2001) allowing researchers to transcend the borders of their original disciplines, creating a "homogenised theory or model pool" (Gibbons et al., 1994, p. 29), a comprehensive framework that is greater than the sum of its parts (Stokols et al., 2003; Wagner et al., 2011).

An alternative approach is proposed by Jan Schmidt (2008) who attempted to disentangle the plurality of non-disciplinary concepts and theories from the perspective of philosophy of science. Focusing on "interdisciplinarity" used as an umbrella term, he argues that interdisciplinary science can be distinguished from monodisciplinary science in four different dimensions:

- an *ontological* dimension: objects of inquiry in interdisciplinary research

- an *epistemological* dimension: concepts and theory used in interdisciplinary research
- a *methodological* dimension: unique methods used in interdisciplinary research
- a *problem* dimension: goals, purposes and motives of interdisciplinary research

This "philosophy of interdisciplinarity" highlights an important fact: manifestations of non-disciplinary research are the result of an *integration* of elements from monodisciplinary research traditions. In particular, *ontological*, *epistemological* or *methodological* "interdisciplinarity" requires respectively the integration of objects and topics, concepts and theory, or methods and practices, of disciplines. This can be seen as the crossing of disciplinary boundaries in the broad overarching *cognitive* dimension. The defining feature of *problem-oriented* interdisciplinarity, as described by Schmidt (2011), is in its focus on problems not defined by disciplines or academia but by society, and can thus be seen as a crossing of boundaries in terms of the social dimension of disciplines. While a secondary effect, the scope of societal problems will require the combination of scientific disciplines and their cognitive elements as per Gibbons et al. (1994) and others as described in the first paragraph of this section.

Notably, *multi-*, *inter-* and *transdisciplinarity* are characterized by an increasing *level of integration* of their disciplinary foundations (Porter et al., 2006; van den Besselaar & Heimeriks, 2001). Returning to the notion of a continuum of scholarship ranging from *mode 1* monodisciplinarity to *mode 2* transdisciplinarity, we see now that such a continuum of scholarship can be described by the extent to which research integrates cognitive elements from different disciplines. This integration ranges from none (monodisciplinary), through limited (multidisciplinary) and extensive (interdisciplinary) to transcending (transdisciplinary).

The notion that research is conducted in a *continuum of disciplinary integration* will serve as the cornerstone for this thesis. Given a set of disciplines and their cognitive content, *disciplinary scientific research* is research conducted within the cognitive boundaries of one *parent discipline*, exhibiting little to no integration of cognitive elements belonging to other disciplines. *Interdisciplinary scientific research*, then, is an umbrella term (following Porter et al., 2006; Wagner et al., 2011) for research that does integrate different *parent disciplines*' cognitive elements to a notable extent. While it can be considered to contain traditional notions of multidisciplinary, interdisciplinary and transdisciplinary research, we will instead refer to a *level of disciplinary integration* or *level of interdisciplinarity* when discussing the extent of integration of parent disciplines.

Important to note is that by defining interdisciplinarity as cognitive disciplinary integration, we limit our scope to scientific outputs of research – namely the cognitive dimensions of publications and similar documents. Heimeriks (2013) describes two more forms of interdisciplinarity besides scientific output, namely interdisciplinarity in the research process and interdisciplinarity in terms of the contribution of non-academic societal dynamics to knowledge production. These express themselves primarily through the individuals and organisations involved in research activity; in other words, the social dimension of disciplines. Our goal is the creation of a classification scheme for publications based on their text, and it is the scientific output that we aim to classify, not the effort leading up to or interactions surrounding the publication.

2.3: Bibliometric mapping and disciplinary classification

Our aim is to find structure in science using publications' abstract and title text data instead the more traditional citation data. It follows that the primary difference between our proposed method and existing methods is the data used to find this structure. The primary contribution that this thesis aims to make is the necessary alteration to current methods to accommodate this new data. Therefore, it

is of paramount importance that we review current methods for structuring and mapping the science space – the field of bibliometrics and the practice of bibliometric mapping – as many of their underlying principles still apply directly to our new proposed method.

Bibliometric mapping of the science system uses network representations at various levels of aggregation. Nodes may represent anything from single publications to sets of journals while edges represent node similarity and may consist of various types of relationships between nodes. In the literature, we find two dominant approaches to constructing maps of the science system. The first involves citation measures, where the edges in the network are based on citation analysis, either by direct citation relations (Publication A cites publication B, so A and B are related), co-citation (A and B are both cited by C, so A and B are related), or bibliographic coupling (A and B both cite C, so A and B are related). The core assumption in this method is that publications sharing a citation relationship, be it direct or indirect, are likely to cover related subject matter. Research using this approach to bibliometric mapping may differ in terms of aggregation and perspective. For instance, Leydesdorff et al. (2013) construct a global map of science based on the Web of Science (WoS) database. They use a top-down perspective in the sense that their network nodes encompass sets of journals assigned to pre-defined WoS subject categories, while their edges are computed based on the aggregated citations among categories. Meanwhile, Waltman and Van Eck (2012) construct a publication-level classification system of science from the bottom up using individual publications as nodes and their direct citation relationships to compute edges. While these approaches work well when it comes to the creation of maps, their subsequent use for analysis or classification is limited to documents within the same database from which the initial sample was retrieved, or at the very least to documents for which proper citation records exist.

In contrast to citation measures, the second approach to bibliometric mapping relies on content words of publications. Content words are typically used to create word maps or term maps of scientific fields which can be used to display relevant topics within that field, how different topics within a field relate to one another or how fields and their relevant topics evolve over time (e.g. Peters & van Raan, 1993a; van Eck, Waltman, Noyons, & Buter, 2010). Further uses include providing context to citation-based clusters of publications (e.g. Waltman & van Eck, 2012). Alternatively, some research uses content words to link publications or sets of publications, but frequently as an addition to citation analysis. For instance, Braam et al. (1991) construct publication "word profiles" to assess the similarity of co-citation clusters. Bruin & Moed (1993) use cognitive words from corporate addresses from *Nature* and *Science* publications to construct a map of research fields.

The direct usage of text data for finding structure in publications has not been explored very thoroughly, although some scholars have suggested it as a possibility. While Glänzel & Schubert (2003) develop a method for the classification of publications to fields and subfields based on pre-defined field categories and a publication's reference literature, they go on to mention the possibility of using cognitive words instead of cited references. In an essay in the book Practicing Interdisciplinarity (Stehr & Weingart, 2000), Van Raan (2000) describes how publications could be related to one another and to disciplines by constructing lists of keywords per discipline and comparing the occurrence of these words in publications. This is not very different from bibliographic coupling in the sense that two publications are related if they share a certain element – in this case the occurrence of a keyword, in the case of bibliographic coupling a reference to a third publication – but in this case, the method is not restricted by the availability of a reference record of those publications, and instead relies on the occurrence of certain terms in their text body. It was this publication in particular that served as the inspiration for this thesis.

The selection of keywords or terms used to link publications is a matter of debate. Possible approaches range from using pre-defined keyword lists compiled by experts, to using machine learning techniques to automatically identify the most relevant terms with the most power to distinguish between disciplines. Some studies use limited word lists or lexical categories depending on their research aims. For instance, Demarest and Sugimoto (2015) use what they call *discourse epistemetrics*, measuring the occurrence of specific socio-epistemic terms and phrases, in order to differentiate between disciplines and classify publications in limited categories. Meanwhile, Waltman and Van Eck (2012) use noun phrases, sequences of words consisting exclusively of nouns and adjectives ending in nouns, to describe the cognitive content of research fields.

3: Theoretical framework

In this section, we will describe the theoretical framework we derived from the previous discussion on disciplinary and interdisciplinary structures in science and bibliometric mapping methods. We will make explicit how these concepts relate to each other, and expand on how disciplines can be characterised using noun phrases as cognitive elements. We will start with an abstract and simplified outline, after which we will discuss practical nuances to this outline, after which we present how we use natural language processing methods to enhance our theoretical framework. This section closes with a summary of our extended theoretical framework as well as an introduction of the research subquestions that flow from our theoretical framework.

3.1: Simple outline

As discussed in the previous chapter, disciplines are distinct from one another in cognitive and social dimensions. Consider now a single cognitive dimension – in our case, the language used within disciplines. The shape of any discipline in this single dimension is determined by that discipline's boundary in that dimension. This boundary separates those cognitive elements – in our case, specific language elements – that belong to the discipline from those that do not. The extent, or shape, of the discipline within this one dimension can then be defined as the collection of cognitive elements which fall within its disciplinary boundary.

From the perspective of bibliometric mapping based on publication data, a discipline can be considered a collection of publications, its features resulting from the combinations of features of the individual publications that together make up the discipline. The cognitive elements contained within these publications thus define the boundary and elements of the discipline. Conversely, the discipline's associated publications' cognitive elements belong within this single discipline and it follows that these publications are *monodisciplinary* as a result of helping *define* the discipline.

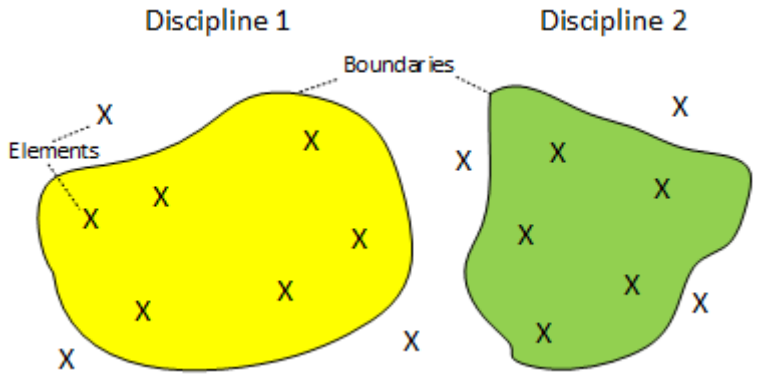


Figure 1. Abstract representation of two disciplines and their cognitive boundaries and elements, in a single dimension.

Now consider a publication not used to define a discipline. Like all publications this one contains a number of cognitive elements and each of these elements may be associated with a certain discipline. Based on the disciplinary association of each of its elements, the publication itself may be more or less associated with various disciplines. The cognitive elements within a publication may be considered as the knowledge that the publication builds on, or the language used in a publication to convey the information contained in it. Given that each of these elements has a parent discipline, the association of the publication as a whole to each discipline may be determined by the elements shared between the discipline and the publication. A publication whose elements all belong to the same discipline is monodisciplinary, while publication whose elements are for the most part associated with a single discipline may borrow elements from a second discipline is multidisciplinary, et cetera. This is where

we return to our notion of interdisciplinarity as a continuum of disciplinary integration. A publication's level of interdisciplinarity is a result of the integration of different disciplinary elements within it, and the disciplines a publication is rooted in may be determined by matching its elements to those of the different disciplines under consideration. See figure 1.

3.2: Practical nuances

The above outline is highly abstract and idealized, and the real world is exceedingly more complex. Since we are working with real data rather than performing a thought experiment, some practical nuances need to be taken into account. First of all, cognitive elements, in our case noun phrases, may not be associated as closely and as exclusively with disciplines as described above. While publications themselves, and the references to or from these, can be considered as representing highly specific pieces of knowledge that can realistically be truly monodisciplinary, language is less well-defined, with words having multiple meanings in different context. The exclusive association of noun phrases to disciplines will be an exception rather than the rule, even more so than that of citation data. A single noun phrase may be associated with various disciplines to various degrees, and as a result its occurrence in a publication may not tell us exactly which discipline this publication is associated with.

Further, the meaning of words is not temporally stable, and the usage of words in communication and publications may wax and wane as language itself is a living thing. Leydesdorff (1997) states that due to this conceptual instability of terms, words alone cannot be used to map the development of science over longer periods of time. As a result, noun phrases as cognitive elements have a limited shelf life and the elements that define a discipline may change over time, as well as noun phrases' association with different disciplines.

On the upside, this fuzzy assignment of noun phrases to disciplines allows us to investigate the similarity in noun phrase occurrence patterns of disciplines themselves. If noun phrases are associated to different degrees with different disciplines, a noun phrase associated strongly with two disciplines forms a cognitive link between these two disciplines. Still, it may also be possible that certain noun phrases appear frequently across disciplines, reducing their power to differentiate between disciplines.

Because of this, and because of computational limitations, many researchers prefer not to work with the entire set of all terms or noun phrases that can be extracted from text, but to select terms with high *relevance* only. Possible approaches range from using pre-defined keyword lists compiled by experts to using machine learning techniques to automatically identify the most relevant terms with the most power to distinguish between disciplines. Some studies use limited word lists or lexical categories depending on their research aims. For instance, Demarest and Sugimoto (2015) use what they call *discourse epistemetrics*, measuring the occurrence of specific socio-epistemic terms and phrases, in order to differentiate between disciplines. In contrast, we, following Waltman and Van Eck (2012), use automatic selection of noun phrases, sequences of words consisting exclusively of nouns and adjectives ending in nouns, to describe the cognitive content of research fields. In the method chapter we will detail the approaches taken to establishing and selecting for the relevance of noun phrases.

3.3: Enhancements from text processing

To account for the fuzzy association of noun phrases to disciplines, we need to move beyond the simple theoretical framework as outlined in the first section of this chapter. To enhance our framework, we turn towards automatic text processing and information retrieval methods. We follow Van Raan (2000), who describes how publications may be interpreted not as terms grouped together, but as

"vectors in a high-dimensional word space" (Van Raan, 2000 p. 74), an interpretation resulting in a *vector space model*.

In this context, words are not mere cognitive elements, but dimensions, their occurrence in a text signifying that text's position in a high-dimensional vector space. Each publication may be represented as a vector denoting the occurrence of terms within its text. These vectors take the form of $\mathbf{d_j} = (t_{1,j}, t_{2,j}, \ldots t_{n,j})$ where $\mathbf{d_j}$ is a document vector for a publication *j* in a discipline with *n* relevant terms, and $t_{i,j}$ equals 0 if term *i* does not occur in the document or otherwise records the amount it does occur. The occurrence of terms in documents gives each document a place in a *vector space* with as dimensions the disciplinary terms. Disciplinary corpora may be represented by combining all the discipline's document vectors into a disciplinary term-document matrix. A mean term occurrence vector, or *centroid*, may then be computed for each discipline. This *centroid* denotes the spatial centre of the document set in our high-dimensional word space, as visualized in figure 2. For a more extensive discussion of vector space models and their applications, I refer to the book *Automatic Text Processing*, Ch.10 (Salton, 1989).

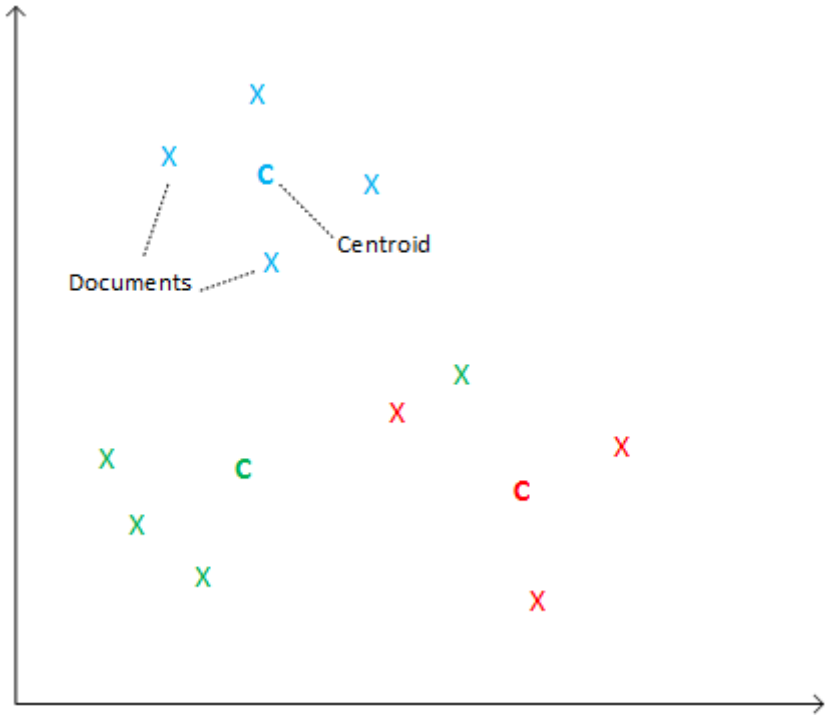


Figure 2: Disciplinary documents make up centroids in a simplified two-dimensional vector space.

The great advantage of enhancing our theoretical framework with the vector space model is that now, publications' cognitive content, in the form of their term occurrence, can elegantly be compared with the overall term occurrence patterns of disciplines by comparing their positions in our high-dimensional vector space.

Figure 3 contains an example of a vector space model. In figure 1, we can see two disciplines **D1** and **D2** each containing four documents, **d1** through **d8**. Each document contains terms, represented by the shapes. Centroids, or average term occurrence vectors, are displayed in the bottom of each discipline, the score per term displayed within the corresponding shapes. A query publication **q** can be matched to these disciplines using some similarity measure – either we compare **q**'s term occurrence with the centroid (curved arrows) or with individual documents within the disciplines directly (best matches displayed with the angled arrows). In either case, the document **q** is more similar to **D1** than

to **D2** – either because it more closely matches the centroid in the former method, or because it more closely matches a document within the disciplines in the latter method.

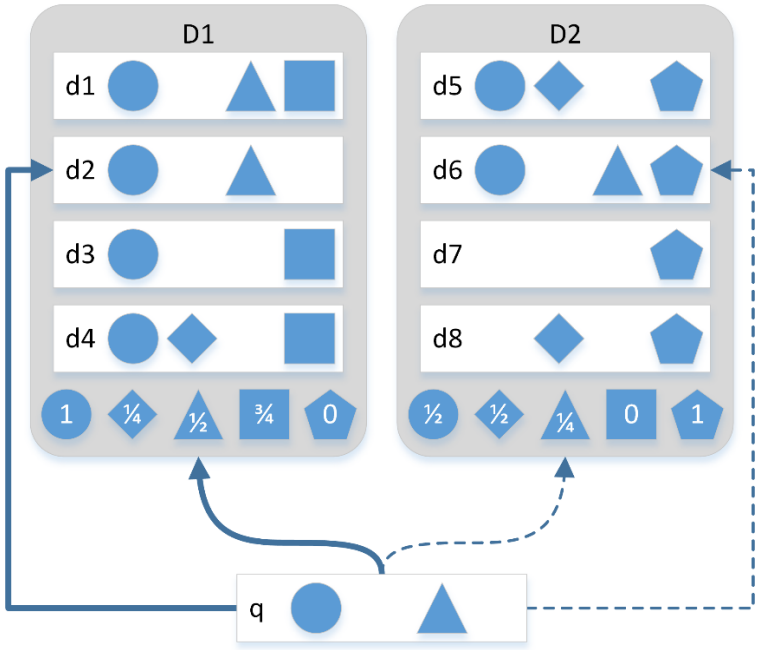


Figure 3: a visual representation of a five-dimensional vector space model containing two disciplines of four documents each and options for matching a query document to those disciplines.

Salton (1989) presents various measures of vector similarity, of which Salton's cosine, or *cosine similarity*, is used frequently in scientometric research (e.g. Boyack, Small, & Klavans, 2013; Leydesdorff & Rafols, 2009; Moed, Glänzel, & Schmoch, 2004; Porter & Rafols, 2009). This measure essentially represents similarity between two document vectors as their multi-dimensional angle. It is a practical choice as it returns easily interpretable similarity scores ranging from 0 (no similarity) to 1 (perfect similarity). The cosine similarity measure does not take into account the magnitude of vectors. This property makes comparisons of centroids and document vectors possible, as matching a sparse, integer-count document vector to a frequency-based centroid vector inherently compares vectors of potentially greatly differing magnitudes. See figure 4 for a visual representation of document similarity comparison using the cosine measure.

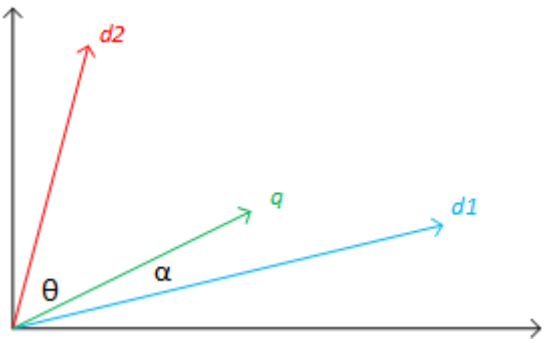


Figure 4: two-dimensional representation of a comparison of a query vector **q** and two document vectors $\mathbf{d_1}$ and $\mathbf{d_2}$. Because the angle α between **q** and $\mathbf{d_1}$ is smaller than the angle θ between **q** and $\mathbf{d_2}$, **q** is most similar to $\mathbf{d_1}$.

3.4: Summary of the framework

Summarizing, we represent disciplines as groups of corresponding monodisciplinary documents. These documents' titles and abstracts contain noun phrases, and the more a noun phrase features within a disciplinary document set, the more associated it is with this discipline. Because these noun phrases represent cognitive elements whose disciplinary association may not be strictly exclusive, it no longer serves to think of disciplines as having strict boundaries which contain their associated cognitive elements.

Instead, we use this fuzzy association of noun phrases to our advantage. If we consider the noun phrases as dimensions, their absence or occurrence in a document denotes that document's position along these dimensions. The document-term vectors used to record the occurrence of noun phrases within documents then become true vectors designating a document's position in a high-dimensional word space.

Groups of monodisciplinary documents, representing disciplines, now occupy a distinct region in this vector space. The position of these disciplines may be defined as the centerpoint – the *centroid* – of their associated monodisciplinary documents. In geometric terms, the centroid is obtained by averaging the position of a discipline's document-term vectors. In less abstract terms, combining the document-term vectors of all documents within a discipline results in a disciplinary document-term matrix, from which the centroid may be obtained by averaging all the values in the term columns across documents.

Each pair of vectors in this vector space model may be assigned some similarity score based on their relative positions in the vector space. For two disciplines, their centroid vectors may be compared to establish how similar these disciplines are. Individual documents may be compared to discipline centroids to find the discipline whose associated noun phrases most closely match the document's content. One may also compare individual document vectors, to find the best match for a query document in a set of documents whose disciplinary association is known.

3.4: Resulting research subquestions

Having defined disciplinarity and interdisciplinarity and with the basis of content-word-based disciplinary classification addressed, it is time to refine the questions this thesis aims to address. First, in order to find structure in the science space and to come to a sensible classification of publications, we need to be able to differentiate between, and establish the links among, disciplines based on their associated noun phrases. The stability over time of these links and differentiations will be included herein.

> *Q1: Can the proposed method using title and abstract noun phrases differentiate between disciplines and find a consistent structure in a network of disciplines?*

Of the disciplines that can be discerned using the method, the most differentiating noun phrases can be listed. Not only will this provide insight into the cognitive content of these disciplines, but the relevance scores may also be used to refine the method, to hopefully come to a better classification of publications and to ease computational stress by reducing the amount of calculations to be performed.

> *Q2: For each discipline, which are its most relevant noun phrases, and can low-relevance noun phrases be removed while preserving the overall disciplinary structure?*

After having found structure in the science system in terms of disciplines and their similarities, the final step of this thesis is to develop a means of classification of publications into this disciplinary structure.

The accuracy of our classification scheme also needs to be verified. This will primarily hinge on the correct classification of known disciplinary publications.

> *Q3: How accurately can disciplinary publications be classified based on the noun phrases appearing in their titles and abstracts?*

In the following chapter, in sections 4.4, 4.5 and 4.6, we will discuss the exact approach taken to come to an answer to these questions, after which we dedicate a results chapter to each of these subquestions in chapters 5, 6 and 7.

4: Method

The goal of this research is to find structure in science, and to develop a method for disciplinary classification, based purely on publications' title and abstract words. If such a method is to be applied in future systematic scientometric and bibliometric research, it needs to be consistent and capable of processing large quantities of publications, use freely available software for the sake of replicability and, ideally, demand no extraordinary computing power. Our method for achieving our goal can be subdivided in five sequential steps:

1. Collection of publication data
2. Definition of disciplines and further data selection
3. Restructuring the publication data
4. Processing disciplinary term-occurrence data
5. Calculation of term relevance scores
6. Classification of a test sample

Each of these steps will be discussed in a section below. The first two steps are aimed at data collection and preparation and will be described in full. The purpose of the third and fourth step is to derive structure from our data, and this chapter will discuss the method we use to do so. The fifth step is aimed at further validating the method, and at demonstrating one of its more practical applications. Concerning these last three steps, we will limit their discussion in this chapter to the approach taken and the transformations and calculations performed on the data in order to achieve results, while the results themselves will be described in detail in the next chapters.

4.1: Data collection

In order to construct sufficiently detailed profiles of noun phrases for a nontrivial amount of disciplines, we need access to a large amount of publications. This data was acquired at the Dutch Centre for Science and Technology Studies (CWTS), where the vast majority of our research was conducted. CWTS maintains two versions of the WoS database: an unaltered version consisting of the data as received by CWTS from Thomson Reuters (WOSDB) and an enhanced version created to facilitate the institute's research efforts (WOSKB). These enhancements include an improved data structure, the inclusion of citation relations between publications, full-text indexing of titles, abstracts and keywords and an improved assignment of journals to subject categories as well as the NOWT classification scheme (NOWT, 2010). This WOSKB database is especially useful for our research as it contains virtually all data necessary for our research – abstract and title noun phrases on a per-publication basis, as well as WoS subject categories of publications' parent journal on which a disciplinary delineation may be based. As the WOSDB and WOSKB databases are periodically updated to include data newly added to the master Thomson Reuters database, it should be noted that data was extracted in March, 2015.

For the period 2000-2010, all publication entries in the database were downloaded using SQL Server Management Studio 2012. This lengthy window of time was chosen to allow for the sample to be split up into several groups of multiple years should this prove necessary. In practice, each single year turned out to contain plenty of information and data for our purposes. For each of these years, several sets of tables were extracted from the WOSKB database, linking individual publication IDs to the following: parent journal, document type, cited and citing document IDs, full text titles and abstracts, and indexed title and abstract noun phrase IDs. Other tables extracted from the database include one linking journal names to WoS subject categories and NOWT categories, and tables linking noun phrase IDs to the textual noun phrases they represent. Several tables were prohibitively large in terms of memory allocation and were split into smaller tables. The entire extracted sample numbered just shy

of 16 million publications. The amount of publication IDs covered in each year ranges from just under 1.2 million in 2001 to around 1.75 million per year[1] in 2010.

4.2: Defining disciplines

When it comes to defining the disciplines that we use to develop our classification scheme, several approaches from literature may be considered. Some studies adopt pre-existing classification categories in a top-down assessment of the science system (Leydesdorff et al., 2013; e.g. Leydesdorff & Rafols, 2009) while others use clustering algorithms to let disciplinary structures emerge from the publication data itself in a bottom-up approach (e.g. Waltman & van Eck, 2012). The former approach can be considered to build on expert knowledge, while the latter approach has the benefit of potentially avoiding artificial, social divides and focussing on the cognitive structure of science alone. While each approach has its merits, our aim is not to investigate the disciplinary divisions of science per se but to test whether terms can be used in disciplinary classification. Building a new disciplinary delineation from the ground up is beyond the scope of this research, and there is indication in literature that content words are not fit for doing so (Leydesdorff, 1997). Hence, the former approach of relying on established categories is most suitable for our purposes. If we can prove the viability of using noun phrases to map structural similarities and dissimilarities in pre-selected disciplines and come to a useful classification, later research may optimize the disciplines themselves to further enhance the structure and the classification results.

In our investigation, we use for our disciplines the subject categories developed by CWTS for the *Science and Technology Indicators 2010* report of the Netherlands Observatory of Science and Technology (NOWT, 2010). These NOWT categories are a tiered grouping system of Web of Science subject categories At the broadest level these consist of 33 disciplinary subject categories, as well as one category for multidisciplinary journals such as *Nature* and *Science*, and one category named *social sciences, interdisciplinary*. The former category contains journals such as *Nature*, *Science* and *PLOS ONE*, but also smaller national and regional journals, which publish articles regardless of disciplinary association. The latter category is comprised of the WoS subject categories *demography*, *social issues*, *biomedical social sciences* and *interdisciplinary social sciences* – an eclectic collection of subjects which does not appear to have a true monodisciplinary character.

We use the first 33 NOWT categories as disciplines, while omitting the latter two categories as they are not monodisciplinary. Since these NOWT disciplinary subject categories are based on the WoS subject categories, they are assigned at the journal level. As a result, all publications in our copy of the WoS database are assigned to one or more of the NOWT disciplines based on the WoS subject categories of the publishing journal.

Due to the possibility of publications being assigned to multiple NOWT disciplines, not all the publications in the sample will be used. For the sake of simplifying our data, only publications assigned to a single discipline will be included in our initial sample. This will allow us to identify a 'core collection' of publications per discipline per year of which the disciplinary association is unambiguous. We further restrict our sample by limiting the included documents to those of the type "article" only (which includes both proper articles and notes as per the classification in the original WoS database), the assumption being that language will differ between, say, scientific articles and review articles published in scientific journals. By limiting our analysis to one document type only, we reduce these expected variations in language use. The sample is further condensed by selecting only those

[1] Overall, the amount of publications increases year-by-year, with the exception of 2000 and 2001. The amount of publications in the database in these years is very close, but slightly lower in 2001, if one does not select for the presence of abstract and title data.

publications with complete title and abstract text and noun phrase index records. Our complete sample of articles from 2000-2010, belonging to only one NOWT category and with complete abstract and title records, ended up containing slightly over 7.1M publications, with a yearly average of 0.65M publications, a minimum of 0.52M articles in 2000 and a maximum of 0.82M in 2010. Publication counts per NOWT category vary. For total publication counts per NOWT category, see table 1. For a complete breakdown of publications per discipline per year, see appendix A.

AGRICULTURE AND FOOD SCIENCE	ASTRONOMY AND ASTROPHYSICS	BASIC LIFE SCIENCES
184337	103480	570240
BASIC MEDICAL SCIENCES	BIOLOGICAL SCIENCES	BIOMEDICAL SCIENCES
43560	295510	490655
CHEMISTRY AND CHEMICAL ENGINEERING	CIVIL ENGINEERING AND CONSTRUCTION	CLINICAL MEDICINE
804692	16843	1553493
COMPUTER SCIENCES	CREATIVE ARTS, CULTURE AND MUSIC	EARTH SCIENCES AND TECHNOLOGY
186099	20687	241922
ECONOMICS AND BUSINESS	EDUCATIONAL SCIENCES	ELECTRICAL ENGINEERING AND TELECOMMUNICATION
89141	36018	157862
ENERGY SCIENCE AND TECHNOLOGY	ENVIRONMENTAL SCIENCES AND TECHNOLOGY	GENERAL AND INDUSTRIAL ENGINEERING
38130	209387	12664
HEALTH SCIENCES	HISTORY, PHILOSOPHY AND RELIGION	INFORMATION AND COMMUNICATION SCIENCES
101891	50461	16819
INSTRUMENTS AND INSTRUMENTATION	LANGUAGE AND LINGUISTICS	LAW AND CRIMINOLOGY
7236	12121	19296
LITERATURE	MANAGEMENT AND PLANNING	MATHEMATICS
13066	20382	227631
MECHANICAL ENGINEERING AND AEROSPACE	PHYSICS AND MATERIALS SCIENCE	POLITICAL SCIENCE AND PUBLIC ADMINISTRATION
108160	1156162	28371
PSYCHOLOGY	SOCIAL AND BEHAVIORAL SCIENCES, INTERDISCIPLINARY	SOCIOLOGY AND ANTHROPOLOGY
129952	10996	30455
STATISTICAL SCIENCES	MULTIDISCIPLINARY JOURNALS	TOTAL
43455	106952	7138126

Table 1: total publications per NOWT category.

4.3: Restructuring data

After having arrived at our final condensed sample, we restructured the data so that it was more easily accessible for further processing. This step and virtually all subsequent steps were performed using the programming language R and its supporting integrated development environment RStudio. For a more detailed discussion of R and its applications, see appendix E.

The starting point of data processing was the lists of publications per discipline per year obtained during the disciplinary definition stage. The mention of a "list" in this context deserves attention – this is an actual, specific data type in R. A list object is a generic vector containing other objects – for instance, single values, vectors, matrices or even other embedded lists. Objects in lists can be given names in addition to values, allowing for the easy creation of tree-like structures in data without compromising the ability to interpret their structure and contents. For each year in the sample period, a list containing for each discipline a vector with all corresponding publication IDs was generated. Names of these list elements correspond to their discipline names. These separate lists based on year may very well be considered elements of a higher-level 'superlist' of the complete sample but computational limitations prevented the creation of such an element because the entire sample is simply too large to fit into memory all at once. Instead, year-based lists containing discipline-specific data can be read into memory from disk as needed, one or two at a time.

Having obtained lists of publication IDs per discipline per year, the next step was to combine this data with the title and abstract noun phrase data obtained from the WOSKB database. The tables linking publication IDs to noun phrase IDs as extracted from the database were very simple in format, containing a column for publication ID, one for noun phrase ID, and a column for the number of occurrences of this publication – noun phrase combination in the sample. While the tables were ordered by publication ID, the fact that publication IDs appeared multiple times – once for each unique noun phrase ID linked to them – rendered these tables difficult to process. As described in the theory section, we would prefer to represent documents as vectors with the vector elements containing all relevant information about the occurrence of terms within those documents. To achieve this, we transformed the vectors of publication IDs per discipline per year into lists, with element names corresponding to each publication ID but undefined element values. Then, a function was written to

crawl through the publication ID – noun phrase ID tables for both titles and abstracts, producing vectors of noun phrase IDs per publication ID and placing those in their corresponding list elements. If a noun phrase ID's occurrence count was higher than one, the noun phrase ID was simply repeated matching the occurrence count.

The end result is a collection of nestled lists, one for each year in the sample. Each of these list contained a further 35 lists, one for each NOWT category, and each of these NOWT category lists contained all the publication noun phrase data vectors associated with its parent discipline in its parent year. Such a nested list structure allows us to easily call upon large chunks of data, or very specific publication data, using either index numbers or list elements' assigned names. A visual representation of this data structure can be found in figure **X**.

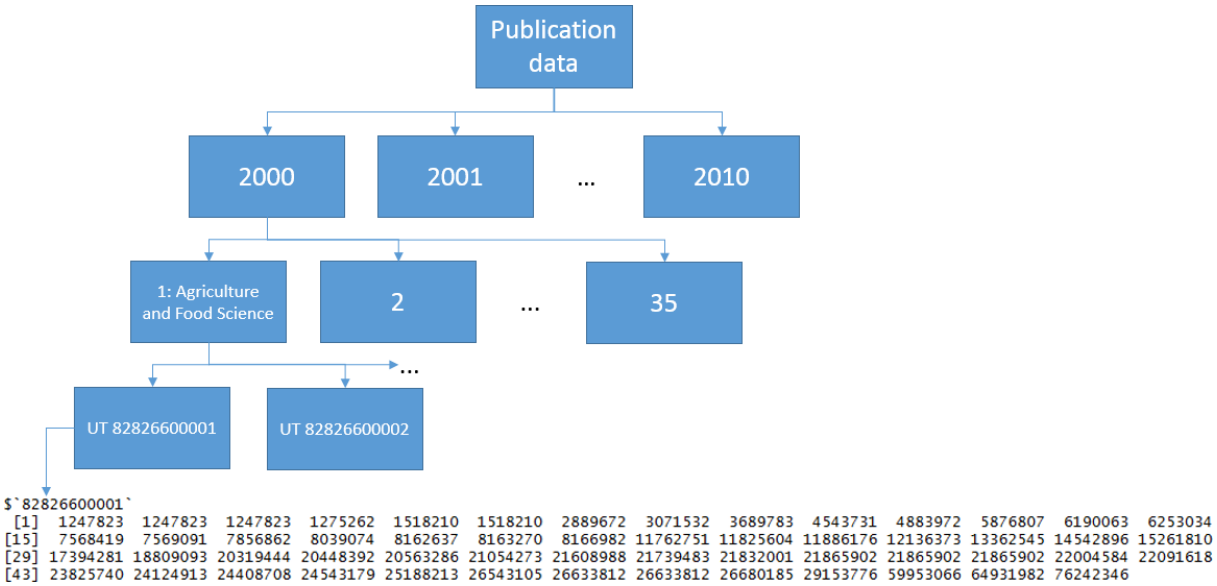


Figure 5: visual representation of our overall data structure.

The reader will notice that the vector representation of documents in our sample differs from the one described in our theoretical framework. Given that the total amount of indexed noun phrases for the WOSKB database is very high (upwards of 110M) while the amount of noun phrases per document is limited by the size of their title and abstract, representing such a sparse vector in full is not practical due to computational limitations. Instead, we save document-term vectors as vectors of noun phrase IDs which feature in said document. This condensed vector representation preserves all relevant information extracted from the database, and is merely a practical matter, functionally identical to the representation in the theoretical framework. This does means that formula such as the cosine similarity do not apply directly to our data as saved on disk, but only to their theoretical "full" vector representation. Even converting our saved data to such a full representation is impractical for significant portions of our sample, and as such internally alternative formulas are used to optimize our computations for our data. The functions used in place of these proper formulae are completely analogous, their differences merely a practical matter and their effect functionally identical to the representation in the theoretical framework.

Similar considerations played a role in deciding the best format for the noun phrase data itself, where we chose to continue the use of noun phrase index numbers rather than transforming them to character-form noun phrases. Saving these as integers is far more memory efficient than saving them as characters, and any operations performed on the data (such as tabulating the total noun phrase occurrence over a set of documents, or finding the intersection of two document vectors) is computationally more efficient when the noun phrases are represented by integers. Tables linking

character-form noun phrase data to noun phrase IDs were extracted from the database during the data collection, and a function was written in our R environment which allowed for extracting character-form noun phrases from these tables.

4.4: Processing disciplinary term-occurrence data

As per our theoretical framework, we mean to describe disciplines' position in the word space as the centerpoint of all documents associated with them. Similar to the document vectors, when calculating the aggregated term occurrence and centroids of disciplines, we depart from the traditional representation of a document-term matrix – a matrix comprised by rows of document vectors and columns representing term occurrence – and instead save our discipline data as lists of document vectors comprised of noun phrase IDs. R has built-in functions to more easily process list data. Using these functions a new total noun phrase occurrence vector can be generated for each discipline in each year, which can subsequently be divided by the amount of documents in the discipline to produce the centroid – the average term occurrence vector – of the disciplinary document vectors. This centroid can be interpreted as the average position of all points designated by the disciplinary document vectors in a multidimensional space, and comparing these centroids between disciplines is a straightforward way of comparing the disciplines as a whole with one another as discussed in the theory section.

Compiling discipline centroids, both for each year in the sample individually and for the complete sample, allows us to compute the cosine similarity of each pair of disciplines as per equation 1.

Equation 1: $$similarity(\mathbf{d}_j, \mathbf{q}) = \frac{\mathbf{d}_j \cdot \mathbf{q}}{\|\mathbf{d}_j\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{n} d_{i,j} q_i}{\sqrt{\sum_{i=1}^{n} d_{i,j}^2} \sqrt{\sum_{i=1}^{n} q_i^2}}$$

In equation 1, $\mathbf{q}$ denotes a query vector and $\mathbf{d}_j$ denotes a document or centroid vector this query is compared to. The cosine similarity returns a value ranging from 0 for completely dissimilar vectors to 1 for completely similar vectors, based on the angle between the vectors and irrespective of the vectors' magnitude. This analysis will allow us to identify which disciplines cover more or less similar terms, and high cosine similarity values will alert us to potential excessive term overlap. For a more detailed discussion of cosine similarity, I refer to Peters & Van Raan (1993b). Details on the implementation of these formula in R and the visualization in VOSViewer can be found in appendix **E**.

In the first subsection of the first results chapter we will present the maps generated using this method, both for the complete sample to display the overall structure, as well as of different single years, to verify whether these structures hold over time. Maps should, however, only be considered a visual aid for discerning structure in networks. After all, they are two-dimensional representations of more complex networks, their shape dependent on a variety of parameters that can be tweaked to one's leisure.

To verify whether the structures uncovered by the maps are supported by the data, as well as similar over time, we calculated the Pearson correlation coefficient of the disciplinary cosine similarity data of each pair of years in the sample. Furthermore, we computed the within-discipline cosine similarity for each discipline over all pairs of years in the sample, to investigate the change or stability of the noun phrases used within each discipline. These results will be presented in the second and third subsection of the first results chapter, respectively.

To close this section of our research, we compare our method of finding structure between disciplines with a more traditional, citation-based method. We had already extracted cited-reference data during our data collection steps, and this data was processed for each of the publications in each discipline in

the sample. The cited references in publications were treated in very much the same way as the title and abstract noun phrase occurrence data, considering them cognitive elements belonging to publications in accordance with our theoretical framework. The result is a measure for similarity in disciplines' cited reference patterns – not unlike bibliographic coupling. To compare the noun phrase based method with the bibliographic coupling based method, a map of the cited reference similarities was made and a Spearman correlation test was performed. We chose Spearman correlation over Pearson correlation as the shape of the distribution of similarities between the two methods may not necessarily match, but overall we expect that, should the methods uncover similar structures, edges that score low using the one method will also score low using the other, and the same goes for high-scoring edges. This follows previous research comparing different similarity measures (Leydesdorff, 2008). The results of this investigation will be presented in the fourth subsection of the first results chapter.

4.5: Calculating term relevance scores

Up until this point we have worked with noun phrase IDs to describe the linkages between publications and disciplines by their noun phrases, irrespective of the meaning of these noun phrases or their relevance to disciplines. While the proposed method allows us to compute a similarity (and conversely, distance or dissimilarity, as 1-similarity) measure between disciplines, we have not yet touched upon which set of noun phrases exactly define these similarities and differences. In other words, which noun phrases contribute most to the identity of disciplines? If one were to describe disciplines by their most relevant noun phrases, which noun phrases would one choose, and why?

In this section, we discuss how we compute disciplinary relevance scores for each term in the sample. These relevance scores serve two purposes: they allow us to select high-relevance terms for each discipline, providing an indication of their cognitive content, and they allow us to prune the term lists of low-relevance terms, easing computational loads in further steps and potentially improving classification in the next step of the method by removing 'noise' terms without any prevailing dominant disciplinary association.

When discerning the relevance of noun phrases, a balance must be found between the absolute frequency of terms and the relative frequency of terms. Terms with a very low absolute frequency are not relevant because compared to other terms in the sample, their occurrence is low, meaning they have little power to describe samples. Terms also need a high relative frequency – comparing their in-discipline frequency to their total frequency in the entire sample – otherwise they cannot be considered distinctive for their discipline.

Waltman and Van Eck (2012), when building their classification system of science, describe the topics of the clusters generated by their method using the most relevant noun phrases for each of these clusters. Their measure of relevance is computed as $n_{ut}/(n_{vt}+m)$ wherein n_{ut} is the amount of publications in field u in which noun phrase t occurs, n_{vt} the amount of publications in field u's parent field v in which noun phrase t occurs, and m a parameter. The intention is that by dividing a field's (or a discipline's) noun phrase occurrence by a parent field (or the total) noun phrase occurrence, one obtains a measure indicative of the relevance of term t to field u. Still, such a measure will have a bias towards low-occurrence terms, which are easily contained within a single discipline. The parameter is added to control for this, giving low-occurrence terms a penalty based on the size of the parameter (25 in the cited publication). This has the advantage of reducing the score of terms which occur infrequently, allowing for a comparison of the relevance of terms within a discipline. It has one major downside, and that is that given disciplines of varying sizes, low-frequency terms will be penalized more severely in small disciplines than in large disciplines if the parameter stays constant. To account

for this effect, we change the way in which we compute the relevance of noun phrases per discipline. To calculate a relevance score, we use equation 2.

Equation 2: $$relevance_n(t, disc) = n_{t,disc}/(n_{t,tot} + m * n_{p,disc})$$

In equation 2, we divide $n_{t,disc}$, the amount of times a term occurs within a discipline, by $n_{t,tot}$, the total term occurrence in the sample, plus a parameter m multiplied by $n_{p,disc}$, the amount of publications in the discipline. While m is still an arbitrary parameter, we found that setting it to 0.001 imposes not too strict a penalty on low-frequency terms but still provides better-interpretable lists of most relevant noun phrases per discipline. The parameter can be considered a tipping point at which frequency a term's relevance score is determined most by its relative or absolute occurrence – 0.001 sets this tipping point at a frequency of 1 term occurrence for every 1000 documents in the discipline. Correcting for the size of disciplines should allow us to compare the relevance of noun phrases not only within, but also between, disciplines. An advantage of this method of relevance calculation is that the relevance of a term will range from 0 to 1, leaving it more easily interpretable.

A downside of the relevance calculation in formula 2, however, is that while we accounted for the discipline size in the denominator, the total amount of term occurrences in the numerator may still be influenced by the noun phrase occurrence in the discipline under consideration. This is necessary for the normalization of the score between 0 and 1, but may still skew results towards larger disciplines dominating the overall term occurrence. To account for these effects, we also introduce a second means of computing term relevance, as displayed in equation 3.

Equation 3: $$relevance_f(t, disc) = f_{t,disc}/(f_{t,tot-disc} + m)$$

In equation 3 we divide the disciplinary term frequency $f_{t,disc}$ by the term frequency in the rest of the sample (total sample minus discipline), denoted as $f_{t,tot-disc}$, plus a parameter m. Similar to equation 2, this parameter was set to 0.001, putting the tipping point at a frequency of 1 term in every 1000 publications. While this equation does a better job at accounting for the effects of large disciplines on the total term occurrence or frequency in the complete sample, a downside of this equation is that the relevance score is no longer bound between 0 and 1.

Both these equations will be used to determine the total noun phrase relevance scores for our disciplines. For each discipline we will extract the 100 most relevant noun phrases, and briefly discuss the results and use them to reflect on the quality of the relevance functions. To close this section of our research, we took each disciplinary centroid from the full dataset and explored how the map structure uncovered by the disciplinary similarities changed as overall low-relevance terms were pruned from the sample.

4.6: Classification of a test sample

The belongingness of a publication to a discipline may also be computed using cosine similarity as described in the theoretical framework. Computing and comparing these per-discipline similarity scores for a query publication allows us to see which discipline's term occurrence patterns match the query publication best. We used this to demonstrate the validity of our method. The publications in our sample should, if our method is correct and our disciplines properly defined, be assigned to the disciplines their NOWT category correspond to – assuming they were assigned to the proper NOWT categories in the database to begin with. Previous research indicates that the allocation of journals to WoS subject categories is not perfect (Leydesdorff et al., 2013), so it is unrealistic to expect perfect classification.

To classify fairly, it is important that we separate the sample in a training sample, which the classification can be based on, and an independent test sample, which will be classified. Because of the possibility of the disciplinary noun phrase profiles changing over time, test sample and classification sample are always extracted from the same year set. For each year, we randomly select 10% of the publications of each discipline to form our test sample, while the remaining 90% form our training sample. These test and training samples are structured in the same way as the overall data structure discussed in section 4.3. The centroids of these training samples were recomputed, as well as their noun phrase relevance scores.

We then explored two different ways of classifying the test sample. The first one simply selected for each publication the discipline whose cosine similarity score was highest based on the discipline's complete centroid. The second method did the same, except it pruned the discipline centroids of noun phrases with overall low relevance scores. These classification methods work in a similar way to the ones described in the theoretical framework, in particular the publication matching methods described in figure **XA** in section 3.3.

For both these classification methods, the centroid similarity and pruned centroid similarity classifications, a script was written which generated an output nested list similar to our data structure, divided first by year, then discipline, then test publication ID. For each of these test publications, the similarity between the document noun phrase occurrence vector and the training set discipline centroids in its corresponding year were computed and combined in a new vector, which was then assigned to its proper place in the output nested list structure. A second script took this data and for each test publication selected the discipline with the highest similarity score, and then computed the classification performance by dividing the amount of correct classifications per discipline by the amount of publications in the test sample discipline.

5: Results – noun phrase occurrence similarity structures

In this section we present the results of our investigation into the similarity patterns uncovered in disciplines' noun phrase occurrence. The methods used to arrive at these results have been discussed in chapter 4.4.

5.1: Between-discipline cosine similarities of the complete sample

As per method section 4.4, our first goal in this research was to map the disciplinary similarities of our sample. By tabulating the noun phrase ID occurrences for each discipline in each year, combining these tables for each discipline, and dividing their values by the total amount of publications in the discipline, we calculated centroid vectors for all 33 disciplinary NOWT categories. These centroid vectors contain the average occurrence of the noun phrases in documents within the discipline. In total, these centroids contain 39,307,186 unique noun phrases and their frequencies of occurrence. For each pair of disciplines, the cosine similarity of their centroid vectors was computed, and a network file was created using these cosine values for processing in VOSViewer. The resulting map can be found in figure 6. Each node represents one of our 33 disciplines. The size of the nodes corresponds to the amount of publications within the discipline. The grey edges between the nodes represent which discipline pairs have the highest similarity scores. Distance between nodes roughly corresponds to decreasing similarity of the discipline pair, but due to the fact that this is a two-dimensional projection of a more complex structure, these distances are not absolute. The labels of each node have been cropped to a maximum of ten characters to improve legibility and to avoid crowding the image.

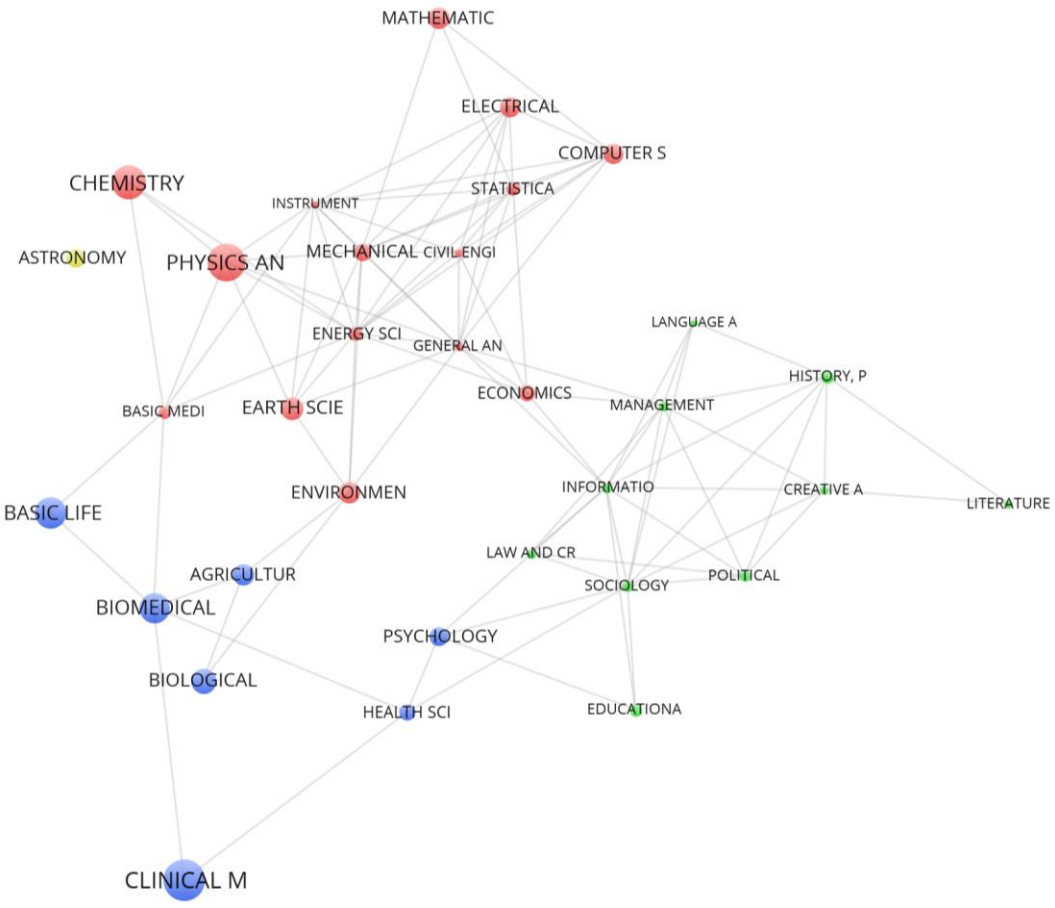


Figure 6: VOSViewer map of discipline similarity network, complete sample.

The software uses several parameters to generate two-dimensional maps from network files, most notably mapping attraction and repulsion, cluster resolution, and normalization. The visualization is further affected by label size, size variation, character length, and the number of lines drawn. Unless specified otherwise, we leave attraction and repulsion values at their default, and use no normalization of the edge values for the mapping parameters. Label length is set to 10, size is set to 1.15 and size variation to 0.3 to avoid cluttering the map, and the 100 strongest edges are drawn. Clustering resolution is varied to produce the most easily identifiable cluster results. The most striking and relatable cluster results were obtained using the maximum clustering resolution still producing four clusters. Node size reflects the amount of publications in the discipline.

Figure 6 already allows for some remarkable observations. We see that the disciplines *clinical medicine*, *literature* and *mathematics* are decidedly at the far edges of the map. The *astronomy and astrophysics* discipline meanwhile appears not to have strong similarity to any other discipline, with all its edges outside the top 100. While several clustering solutions are possible based on the chosen clustering resolution parameter, the most striking and intuitive clustering solution occurs when choosing the parameter so that no more than four clusters appear. In this clustering solution[2], four clusters appear: a STEM[3] cluster (red), a life sciences cluster (blue), a humanities cluster (green), and the lone astronomy and astrophysics discipline cluster (yellow). These clusters include the following disciplines:

- STEM cluster (red)
    - Basic medical sciences
    - Chemistry and chemical engineering
    - Civil engineering and construction
    - Computer sciences
    - Earth sciences and technology
    - Economics and business
    - Electrical engineering and telecommunication
    - Environmental sciences and technology
    - Energy science and technology
    - General and industrial engineering
    - Instruments and instrumentation
    - Mathematics
    - Mechanical engineering and aerospace
    - Physics and material science
    - Statistical sciences
- life sciences cluster (blue)
    - Agriculture and food science
    - Basic life sciences
    - Biological sciences
    - Biomedical sciences
    - Clinical medicine
    - Health sciences

[2] In all our noun phrase based VOSViewer maps including the ones in section 5.2 and appendix **B**, there is a clustering resolution which results in four clusters, one made of only *astronomy and astrophysics* and the others as described in the list. Because of this consistent feature, we chose to consistently opt for a clustering resolution which produced this distinct four-cluster solution. Clustering resolutions resulting in more than four clusters quickly devolve into difficult to interpret structures.

[3] An acronym introduced and used by the United States National Science Foundation (see, for instance, National Science Board, 2014) for Science, Technology, Engineering and Mathematics. “Science” refers to the fundamental sciences. The term is roughly equivalent to the Dutch *bètawetenschappen*.

    - Psychology
- Humanities cluster (green)
    - Creative arts, culture and music
    - Educational sciences
    - History, philosophy and religion
    - Information and communication sciences
    - Language and linguistics
    - Law and criminology
    - Literature
    - Management and planning
    - Political science and public administration
    - Sociology
- Astronomy and astrophysics (yellow)

Other features also make intuitive sense: *economics and business* bridges the humanities and STEM clusters, and *environmental sciences and technology* and *agriculture and food science*, as well as *basic medical sciences* and *basic life sciences* appear prominently at the border of the STEM and life sciences cluster. Psychology meanwhile lies on the border of the humanities and life sciences clusters as one might expect.

However, the map comes with several caveats. First, it is but a two-dimensional representation of a network resulting from similarities over many thousands of dimensions, and the current projection is a result of several parameters whose values greatly influence the shape of the map. It is therefore not an ideal representation, in fact, there is no ideal representation in two-dimensional space. It is but one of many different possible visualizations of the discipline similarity network. It is important not to put too much stock into observations made using only the map, especially considering that it is human nature to perceive patterns even in meaningless information.

Second, the map is a result of the similarity of the aggregated disciplinary publications from all years in the sample. It tells us nothing about the stability of the between-discipline patterns and clusters over time, or the internal stability of the disciplinary noun phrase occurrence patters. Just by looking at the map we are unable to determine whether we have found true and stable structure or if the current projection is simply the result of an averaging of wild and unpredictably changing patterns.

5.2: Between-discipline cosine similarities of sample year segments

In order to address these pitfalls, we delved deeper into the data. First, we used R and VOSViewer to generate similar maps for all individual years in the sample. A selection of these maps are displayed in figure 7, while the entire collection can be found in appendix B. VOSViewer produces maps with varying orientations, and as the orientation of the maps is essentially meaningless, the maps were mirrored and rotated to match the complete sample map in figure 6.

Visually these maps show strikingly similar features. Clustering solutions using again the maximal clustering resolution resulting in four clusters produce the exact same clusters as displayed in figure 6. The disciplines seem to mostly retain their overall and relative positions. The most notable discrepancy between the maps in 7 and the overall sample map is the position of *astronomy and astrophysics*, which trades places with *chemistry and chemical engineering*. A similar swap occurs in figure 7 with *biological* and *biomedical sciences* between the years 2004 and 2007. The mapping procedure, in essence, computes an optimal solution using mapping attraction and repulsion forces based on the network edge values. Small changes in the between-discipline similarity scores may tip the balance of these forces one way or another, resulting in a new optimal solution and, visually, a positional swap of

network nodes. Thus these swaps are likely the result of the two dimensional projection algorithm. Still, the underlying similarity scores deserve closer scrutiny.

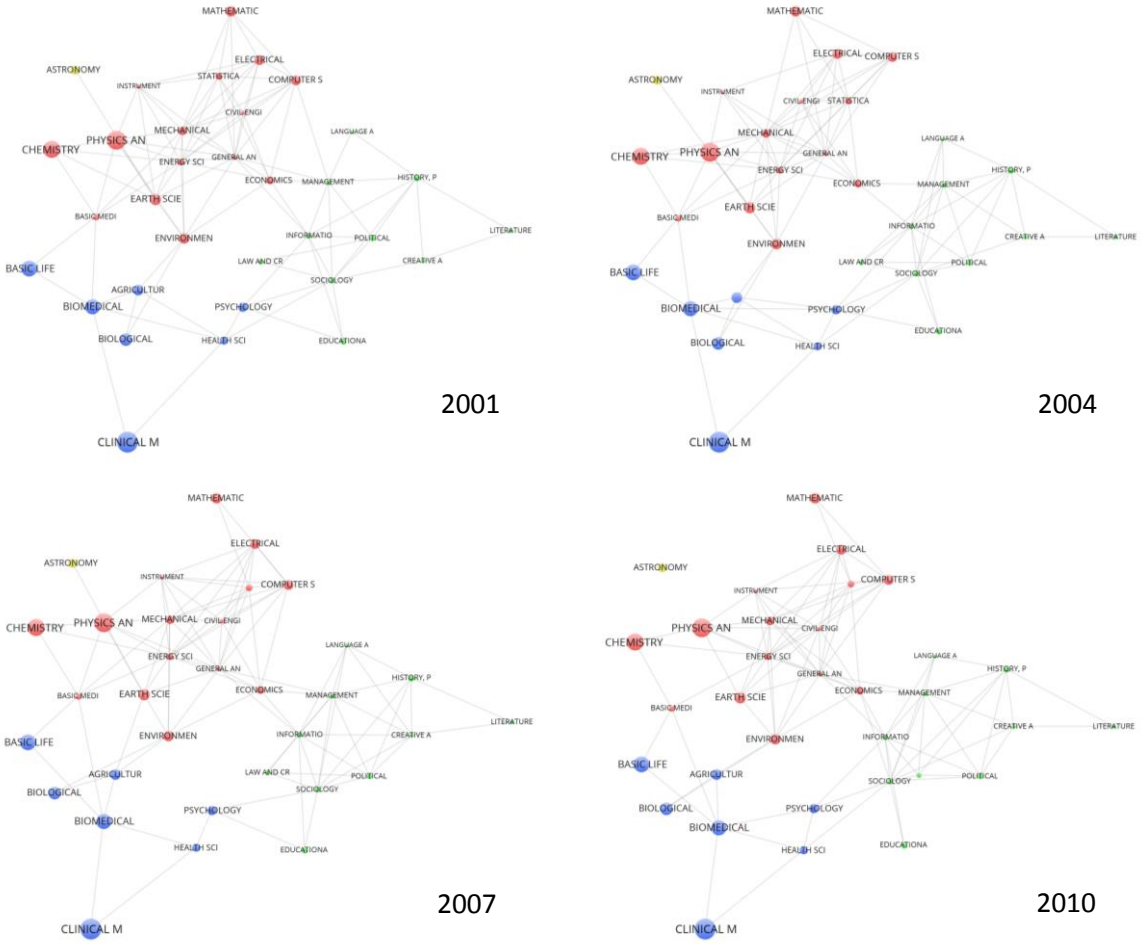


Figure 7: VOSViewer maps of the discipline similarity network, single-year data segments.

While the overall structures appear to be very constant, the information contained within these maps is limited. To definitively establish whether the structures are constant over the years in the sample, and to rule out that any differences are the result of large changes in discipline pairs' similarity scores, we plotted the subsequent years' cosine scores against each other and tested their mutual association using Pearson's product-moment correlation. These plots and the results of the correlation tests may be found in appendix C. A selection of these has been included in figure 8.

In each of the plots in figure 8, we plot all discipline pair cosine similarity scores from one year against the same scores for a subsequent year. Each point in the plot can be considered to represent one of the edges in the discipline similarity networks, its position determined by its value in the two years. As we can see, the cosine similarities between discipline pairs from subsequent years are highly correlated. Squaring the correlation coefficient gives us the explained variance, which in all subsequent year tests lies above 98%. This confirms our observation that the discipline noun phrase similarity structure is remarkably stable. To determine whether this stability holds over larger time periods, we perform the same correlation test on more distant year samples. The results for the comparison between the 2000 and 2010 data can be found in figure 9, while all the correlation coefficients for each pair of years are plotted in figure 10.

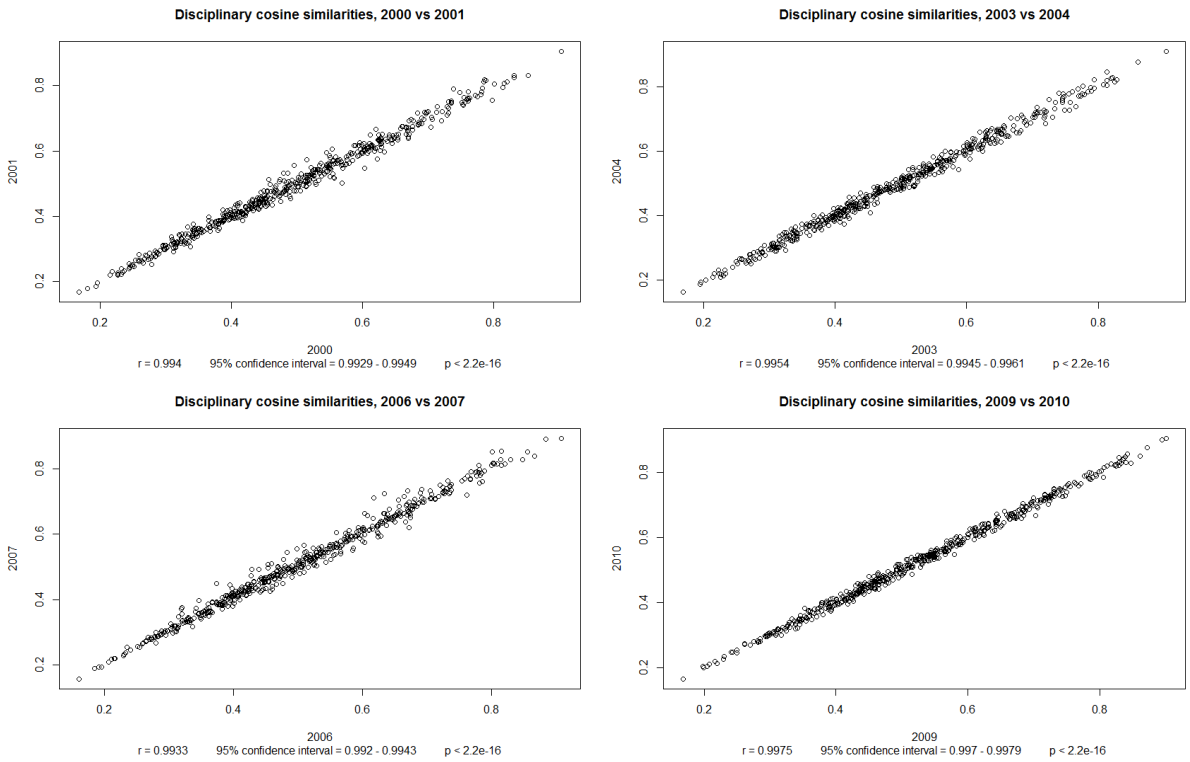


Figure 8: plots and correlation test results of selected subsequent year samples.

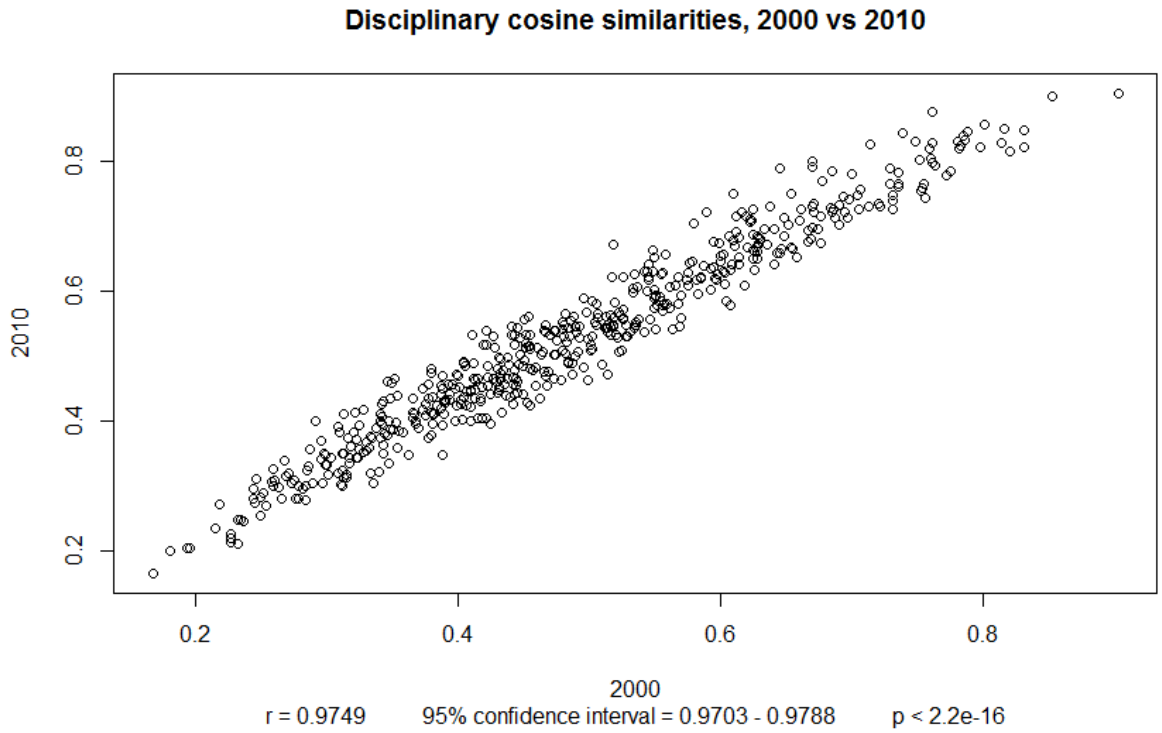


Figure 9: plot and Pearson correlation test results of 2000 and 2010 data.

Figure 9 shows that the cosine similarities from 2000 and 2010 differ more strongly than those of subsequent years. This implies that while the discipline noun phrase occurrence patterns are highly similar when comparing any two subsequent year pairs, the small differences 'add up' to lead to larger differences in more distant sample segments. This may seem like a trivial observation, after all, the fact that new discoveries are made in science implies that over the years new subject matter will be

discussed, but it does confirm our expectation that the occurrence patterns are not stable. When plotting the correlation coefficient of all year-pairs in figure 10, we can see that there is definitely a downwards trend in correlation coefficient as the amount of years between two sample segments increases. This implies that the language used in our disciplines does indeed change over time and that these changes are cumulative or otherwise directional or evolutionary, rather than a random chance variation around some static average disciplinary term occurrence vector. Correlation coefficients remain high even for more distant samples, with explained variance barely dropping below 95%.

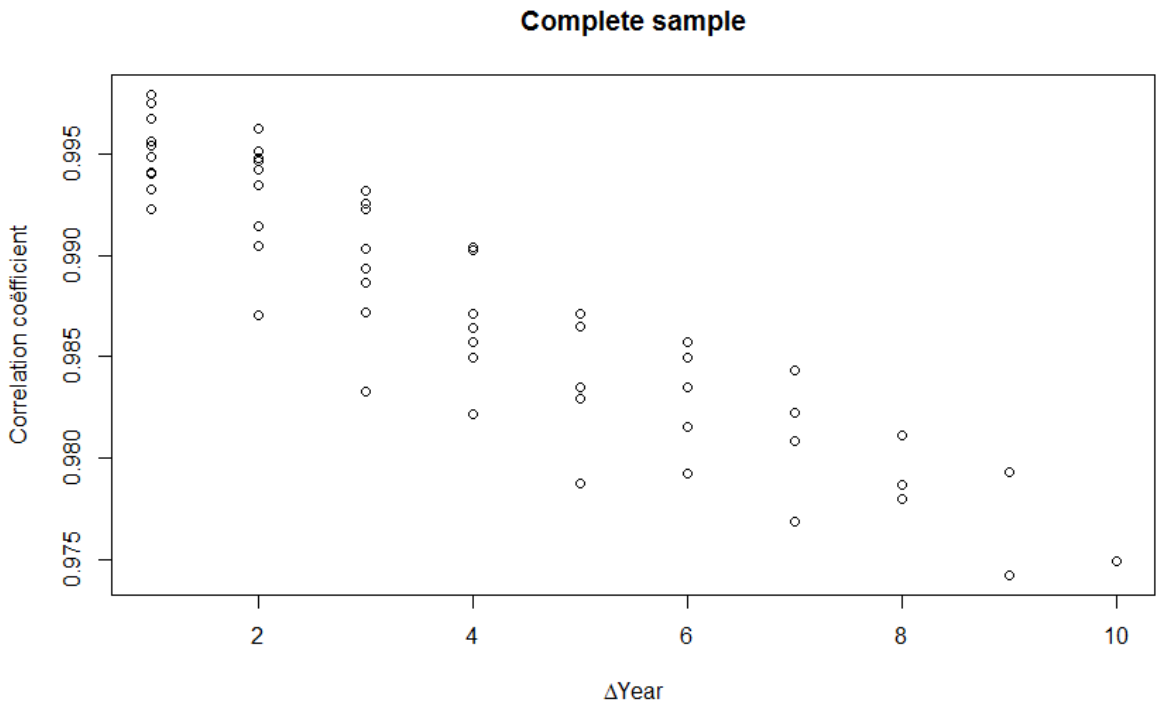


Figure 10: correlation coefficients of each pair of years in the sample. The size of the year-gap between samples is displayed on the x axis, ranging from 1 (including 2000-2001, 2001-2002, etc.) to 10 (2000-2010).

5.3: Within-discipline cosine similarity

To investigate this phenomenon further, we computed within-discipline cosine similarities of their centroid vectors over the years. Selected results are displayed in figure 11 while the entire collection of plots can be found in appendix D.

While the downwards trend in similarity by sample year gap is also evident here, this pattern seems more pronounced in some disciplines, and more erratic in others. While, for instance, the within-discipline centroid cosine similarities for *clinical medicine* seem to decrease neatly as the year gap between samples widens, *mechanical engineering and aerospace* behaves not quite as neatly, with overall high similarity scores for certain pairs of year samples and lower similarity scores for others. Judging purely from the plot, it seems as though the change in noun phrase occurrence patterns remains fairly stable from year to year for a certain period of time, then suddenly a large change occurs, after which the rate of change goes back to the original pace. The simplest explanation for this phenomenon may be the indexation of a new journal in the WoS database, but it could also be a sudden disruptive development in the field which causes a shift in focus. While we cannot rule out either possibility definitively, when we look at the complete sample data in appendix **A**, we see a sharp increase in the amount of publications in the *mechanical engineering and aerospace* discipline

between the years 2005 and 2006. This coincides roughly with the single high-similarity record with a year gap of 6 in figure 11, suggesting that this is indeed a result of newly indexed publications.

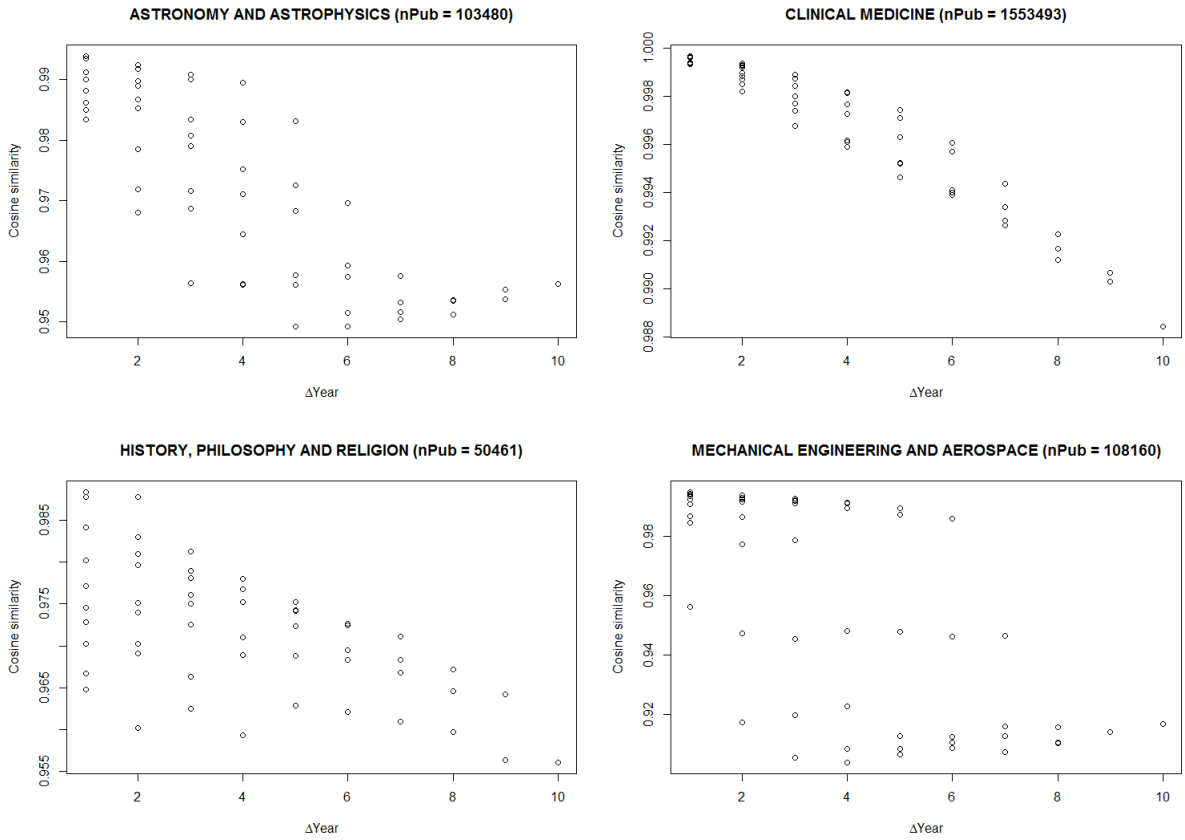


Figure 11: selected within-discipline centroid cosine similarities over the years.

Unfortunately the details of these plots remain difficult to interpret. What one should take away from this data is that overall it is in line with our previous observations: similarities between samples decrease as the amount of years between two sample segments increases, even though both remain exceptionally high. The noun phrase occurrence patterns are therefore highly stable over time.

6: Results – disciplinary noun phrase relevance

In this chapter we discuss the results obtained from the relevance scoring of noun phrases. In the first subsection we describe lists of high-relevance noun phrases for each discipline in the complete sample and their implications as to which approach to establishing relevance is superior. In the second subsection we use both approaches to enhance our mapping method by pruning low-relevance noun phrases. The aim is both to hopefully reduce a portion of discipline-pair similarity caused by 'noise' induced by similarities in inconsequential terms, but also to reduce computational demands of our mapping and classification method. The original map in figure 6 in section 5.1 was constructed using the disciplinary occurrence data of over 39 million unique noun phrases, and while using this complete data set for classification is not impossible, it would be much more efficient if we could remove low-relevance terms while preserving the overall structure of the map to achieve an equivalent result faster.

6.1: Most relevant noun phrases per discipline and a comparison of relevance scores

As addressed in the methods section, there are several ways to establish which noun phrases are most associated with particular disciplines. In this section we present the results of two of these approaches. The first is an adjusted form of the relevance score used by Waltman & Van Eck (2012), based on absolute disciplinary and total occurrence of noun phrases as detailed in equation 2 in section 4.5. The second uses the relative frequency of noun phrases in the discipline and in the rest of the sample, as detailed in equation 3 in the same section. We refer to these two approaches as the *occurrence relevance* and the *frequency relevance*, respectively.

Both these equations include a parameter to penalize low-frequency and low-occurrence terms. This parameter was set to 0.001. As noted by Waltman & Van Eck (2012), these types of parameters are somewhat arbitrary, and their chosen value was a result of trying out different values to see which yielded satisfactory results. More on this topic will follow in the discussion chapter.

For each of these approaches, the relevance scores of all noun phrases in the complete sample for each discipline was computed. The top 10 relevant noun phrases for both methods for a selected few disciplines are displayed in table 2.

These results reveal some interesting features of the two relevance measures. First, the occurrence relevance seems to favour composite noun phrases, comprised of combinations of words, while frequency relevance seems to have more single-word noun phrases in its top relevant lists. Second, the occurrence relevance measure seems to favour more obscure terms, while the frequency relevance measure seems to contain a lot more general terms. In *basic medical sciences*, occurrence relevance seems to produce questionable results – we will discuss these at the end of the subsection.

These differences in the relevance results can be explained by the details of the equations that were used to compute relevance. For simplicity's sake, let us assume for a moment that the parameter m in equations 2 and 3 were set to 0. To compute the occurrence relevance we use equation 2, dividing the disciplinary noun phrase occurrence by the total noun phrase occurrence. This means that, not only is the occurrence relevance score bound between 0 and 1, the occurrence relevance for each term sums to 1 across disciplines. This effectively means that terms with a high occurrence relevance in one discipline need to have low occurrence relevance scores in other disciplines, and the occurrence relevance favours terms which are *uniquely relevant* to single disciplines. If, as in practice, m is not set to zero, the total occurrence relevance of a term instead sums to a value less than one, but the point that the equation favours uniquely relevant terms still stands.

Physics and material science

Occurrence relevance	Frequency relevance
physics	physics
american institute	american institute
optical society	optical society
alloy	alloy
gaas	america
phys	microstructure
america	thin film
gan	phys
superconductivity	mev
grain boundary	film

Literature

Occurrence relevance	Frequency relevance
kosovel	poetry
charlotte bronte	poem
jane eyre	essay
trollope	text
baudelaire	poet
mallarme	fiction
balzac	writer
kleist	writing
zola	shakespeare
quevedo	genre

Clinical medicine

Occurrence relevance	Frequency relevance
surgery	surgery
resection	patient
conclusions	complication
complication	recurrence
overall survival	consecutive patient
recurrence	chemotherapy
surgeon	diagnosis
visual acuity	risk factor
consecutive patient	case report
surgical treatment	dog

Astronomy and astrophysics

Occurrence relevance	Frequency relevance
m circle dot	galaxy
ngc	star
star formation	ngc
galaxy	m circle dot
kpc	kpc
stellar population	star formation
h ii region	luminosity
stellar mass	redshift
globular cluster	planet
early type galaxy	metallicity

Basic medical sciences

Occurrence relevance	Frequency relevance
inc j biomed mater res	scaffold
inc j biomed mater res part b	biocompatibility
biomedical engineering society	biomaterial
inc j biomed mater res 92a	tissue engineering
inc j biomed mater res 93a	structure activity relationship
appl biomater	hydrogel
inc j biomed mater res part a 95a	sbf
inc j biomed mater res 91a	body fluid
inc j biomed mater res 90a	plga
appl biomater 90b	wiley periodical

Table 2: the ten most relevant noun phrases in five disciplines, according to both the occurrence relevance method and the frequency relevance method.

In contrast, equation 3 divides the frequency of a term within a discipline by the frequency of the term in the rest of the sample. This means that the numerator and denominator are independent of one another, and that the frequency relevance score is not bound between any values. A high frequency relevance of a term in one discipline does not necessarily preclude it having a similarly high frequency relevance in another discipline, as long as the term frequency in both these disciplines is sufficiently high compared to the term frequency in the rest of the sample. This means that frequency relevance does not favour uniquely relevant terms as much as occurrence relevance, though it may still select them.

This explains the differences we can observe in the top terms selected by each method. Occurrence relevance favours composite noun phrases and other highly specific terms as these are less likely to occur in many different disciplines, while frequency relevance does not shy away from selecting more general terms. This is illustrated nicely by the top terms of the literature discipline. Occurrence relevance selects predominantly writer names, which are very unlikely to occur with any regularity outside this discipline, while frequency relevance gives the highest scores to terms which are more general but still clearly and are intuitively associated with literature, such as *text*, *essay* and *writing*. It is likely that such terms occur with some regularity outside the literature discipline, but they are clearly far more frequent here than elsewhere.

It is clear then that frequency relevance may be superior to occurrence relevance for the purpose of pruning low-relevance noun phrases. Where cosine similarity connects disciplines based on their common noun phrase occurrence patterns, occurrence relevance awards lower scores to these "bridging" noun phrases, instead favouring uniquely relevant noun phrases which are not useful for finding structure at all. Frequency relevance on the other hand may preserve these bridging noun phrases if they are sufficiently relevant to each discipline, while still assigning lower scores to terms which are universally frequent across disciplines. Furthermore, since occurrence relevance compares disciplinary noun phrase occurrence to total noun phrase occurrence, discipline size might skew the results, even though we attempted to correct for this in the use of the low-occurrence penalty parameter. If a noun phrase occurs just as frequently in a small discipline as it does in a large discipline, equation 2 will assign the noun phrase a higher relevance score for the larger discipline simply because this large discipline has a larger share of the total occurrence.

The top occurrence relevance terms of *basic medical science* reveal that this selection of uniquely relevant noun phrases does not always produce results which are as easily interpretable as in the *literature* discipline. This particular result may be attributed to a flaw in the way the noun phrase data was collected. It is clear that if we want to use occurrence relevance to properly convey the unique cognitive content of discipline to a human audience, the noun phrase selection needs to be adjusted. It is a reminder that not all data which may readily be interpreted by computer algorithms is suitable for human consumption.

6.2: Optimizing the mapping process by pruning low-relevance terms

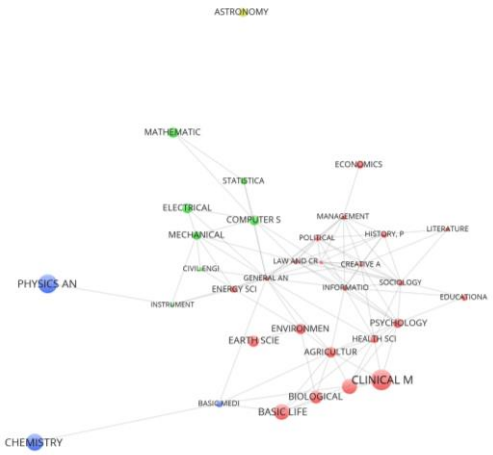
Because of the reasons outlined in the previous subsection, we can already expect frequency relevance to prove to be superior to occurrence relevance when it comes to preserving links between disciplines. We put both methods to the test by pruning the disciplinary centroids vectors of low-relevance terms and using the resulting pruned centroids to generate new discipline similarity maps.

A naive approach to the pruning process would be to take each discipline's centroid vector and to remove all noun phrases whose relevance score lies below a certain threshold. A great downside of this approach is that, since a term has a relevance score for each discipline, terms may be removed from one discipline while being retained in others. This effectively removes these terms' effects on the similarity scores of discipline pairs whenever the term is pruned from one of these disciplines' centroid, leading to an incorrect similarity score.

Instead our pruning process has two steps. First, we wrote a script which selected all noun phrases which scored above our threshold relevance in any discipline and place them in a list of high-relevance terms. Second, for each discipline centroid, all noun phrases that do not occur in this high-relevance term list are pruned. This ensures that noun phrases with high relevance for one discipline are not pruned from other disciplines.

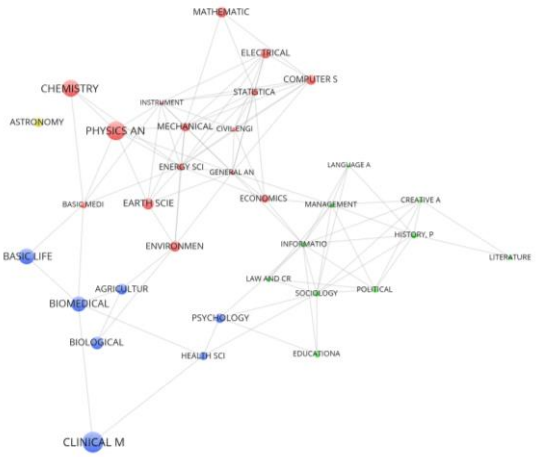
For our pruning thresholds, we used 0.5 and 0.2 for occurrence relevance, and 1 and 2 for frequency relevance. We recomputed the discipline similarity scores and used these to generate new maps with VOSViewer. The results are displayed in figure 12.

Occurrence relevance pruning

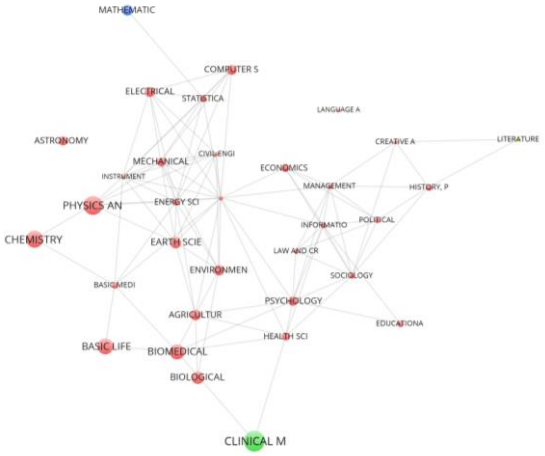


Threshold = 0.5, nterms = 12800

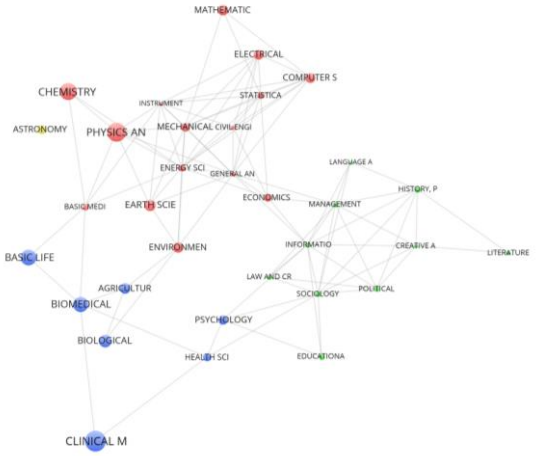
Frequency relevance pruning



Threshold = 2, nterms = 20114



Threshold = 0.2, nterms = 140598



Threshold = 1, nterms = 46482

Figure 12: maps after pruning of low-relevance noun phrases, complete sample.

Comparing the maps in figure 12 with the original map in figure 6 in section 5.1, we see clearly that when pruning based on occurrence relevance, the original structure breaks down, both in terms of positioning of nodes, edge values, and cluster assignment. Frequency relevance pruning, on the other hand, preserves both the overall structure and edges as well as the clustering solution. Choosing a frequency relevance threshold of 2 leaves us with 20114 relevant unique noun phrases, a great reduction from our previous 39 million unique noun phrases.

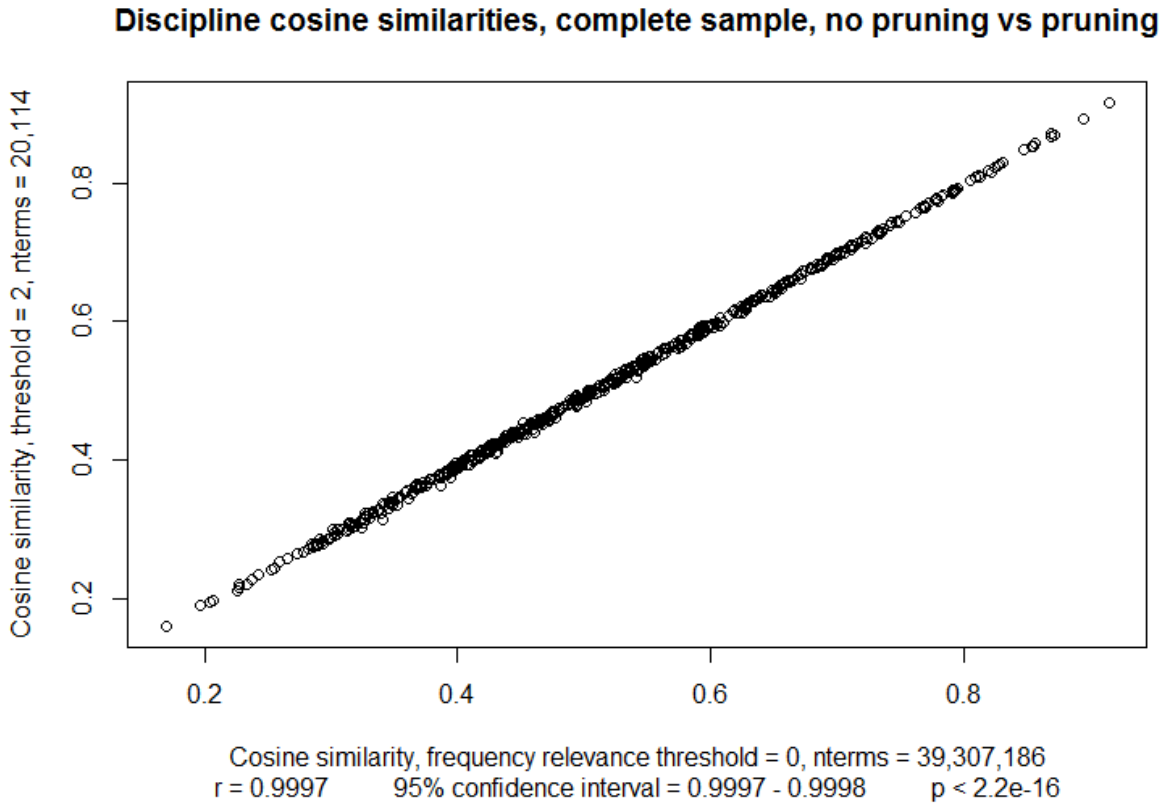


Figure 13: comparison of discipline similarities, no pruning vs frequency relevance pruning

Figure 13 shows a comparison and correlation test of the original between-discipline cosine similarities for the complete sample without pruning as obtained in section 5.1, with the between-discipline cosine similarities after frequency relevance pruning with a threshold of 2, used to produce the top-right map in figure 12. Judging by the linearity of the plot and the high correlation coefficient, the two methods produce an almost identical discipline similarity network. The method with pruning reduces the amount of unique noun phrases used to arrive at this result by three orders of magnitude. This means that only a small minority of the unique noun phrases used in scientific publications are discipline-specific while the vast majority can be disregarded without compromising the structure derived from those publications.

7: Results – disciplinary classification of publications

So far we have demonstrated the existence of relatively stable between-discipline similarity structures, as well as the power of frequency relevance pruning to reduce the amount of noun phrases used to arrive at these structures. In this chapter, we explore how accurately publications can be classified into disciplines, and whether this accuracy increases or decreases after pruning low-relevance noun phrases. First we discuss the construction of our test and training samples and give an example of how our classification process works. Then we discuss the outcomes of the classification for three scenarios: similar-centroid classification without pruning, similar-centroid classification with frequency relevance threshold pruning, and finally direct publication similarity.

7.1: Test and training sample construction and classification example

The selection of the test and training samples was done as described in the methods chapter, section 4.6. For each year in the sample, two test-training sample pairs were generated using random seeds to enlarge the amount of publications that we could classify, in order to increase the reliability of our findings. In total, the combined test samples consisted of 1,404,052 publications: two random draws of 10% of the original sample of roughly 7 million publications. Each test sample has its own associated training sample from which we construct discipline centroids and relevance tables, which is the reason multiple random draws to enlarge the total test sample are possible.

The way the classification proceeded was as follows. First, because the training samples differ from the samples used in the analysis of the between-discipline similarities as described in section 4.4 and chapter 5, their discipline centroids and relevance data had to be recomputed. Then, each publication in the test sample was compared to the (pruned) discipline centroids, and the test publications assigned to the most similar discipline. After classification, the classification performance was computed for each discipline in the sample by combining the classification records for each of the two seeds' test-training sample pairs and all years in the sample.

An example: the very first publication in our test samples was an *agriculture and food science* publication by Altan et al. (2000) titled *"Effects of short-term fasting and midnight lighting on egg production traits of laying hens during summer season"*. In our data, this publication is simply recorded as a vector of noun phrase IDs, as seen in figure 14. Tabulating this data (and, for the purpose of this example, replacing the noun phrase IDs with their corresponding noun phrases) produces the noun phrase occurrence vector in figure 15.

```
$`86634400007`
 [1]   447320   734042  2651782  3053067  3273716  4703086  5197079  5899460  7328324  7328324  7385037  7385037
[13]  7385037  7386238  7390720  7394056  7403525  8615265  8621782 10313143 10599747 10599747 13840322 13859609
[25] 15465995 15465995 15471023 20284235 21469213 21865902 22998084 22999117 22999117 23078894 24232011 24254537
[37] 24254537 24254537 24379615 24585821 24585821 24627315 27359206 61276876 70398496 79159360
```

Figure 14: Data for Altan et al. (2000) in our database.

ad libitum	aim	bird	brown layer	cage
1	1	1	1	1
commercial strain	control	day	effect	egg production
1	1	1	2	3
eggshell quality	egg weight	egg quality	egg production trait	feeding
1	1	1	1	1
feed	h light	hen	lighting period	lighting regimen
1	1	2	1	1
midnight lighting	midnight	present study	rectal temperature	result
2	1	1	1	1
short term	short term fasting	significant decrease	stock	strain
1	2	1	1	3
study	summer season	supplementary lighting	withdrawal	regime bird
1	2	1	1	1
fasted regime	hottest hour			
1	1			

Figure 15: Noun phrase occurrence vector of Altan et al. (2000).

This noun phrase occurrence vector can be matched against the disciplinary centroid vectors in the corresponding training sample using cosine similarity. Doing so for each discipline and combining this similarity data produces the query-centroid similarity vector displayed in figure 16. Our classification scripts simply classify the publication into the most similar discipline. Notice that in this example the highest-scoring discipline is *agriculture and food science*, which, according to our data, is the correct discipline for Altan et al. (2000). This publication has, therefore, been classified correctly.

```
$`86634400007`
                                AGRICULTURE AND FOOD SCIENCE                                          0.22639147
                                                ASTRONOMY AND ASTROPHYSICS                            0.08212497
                                                                BASIC LIFE SCIENCES                   0.14729369
                                                BASIC MEDICAL SCIENCES                                0.14725870
                                                                BIOLOGICAL SCIENCES                   0.15612480
                                                                BIOMEDICAL SCIENCES                   0.18312100
                                CHEMISTRY AND CHEMICAL ENGINEERING                                    0.08980737
                                                CIVIL ENGINEERING AND CONSTRUCTION                    0.10177085
                                                                CLINICAL MEDICINE                     0.10534449
                                                COMPUTER SCIENCES                                     0.04011912
                                                CREATIVE ARTS, CULTURE AND MUSIC                      0.05034891
                                                                EARTH SCIENCES AND TECHNOLOGY         0.10266877
                                                ECONOMICS AND BUSINESS                                0.10665702
                                                EDUCATIONAL SCIENCES ELECTRICAL ENGINEERING AND TELECOMMUNICATION                    0.09186137                                                0.06792281
                                ENERGY SCIENCE AND TECHNOLOGY                                         0.11583442
                                                ENVIRONMENTAL SCIENCES AND TECHNOLOGY                 0.14917071
                                                GENERAL AND INDUSTRIAL ENGINEERING                    0.12503642
                                                HEALTH SCIENCES                                       0.14238954
                                                HISTORY, PHILOSOPHY AND RELIGION                      0.04054856
                                                INFORMATION AND COMMUNICATION SCIENCES                0.09631953
                                INSTRUMENTS AND INSTRUMENTATION                                       0.10209041
                                                LANGUAGE AND LINGUISTICS                              0.06537984
                                                                LAW AND CRIMINOLOGY                   0.09059220
                                                LITERATURE                                            0.03158360
                                                MANAGEMENT AND PLANNING                               0.06197500
                                                                MATHEMATICS                           0.04890942
                                MECHANICAL ENGINEERING AND AEROSPACE                                  0.14209489
                                                PHYSICS AND MATERIALS SCIENCE POLITICAL SCIENCE AND PUBLIC ADMINISTRATION                    0.12354076                                                0.07020867
                                                PSYCHOLOGY                                            0.15755345
                                                SOCIOLOGY AND ANTHROPOLOGY                            0.08477026
                                                                STATISTICAL SCIENCES                  0.04979009
```

Figure 16: query-centroid cosine similarity vector for Altan et al. (2000) and the disciplines in its corresponding training sample.

7.2: Similar-centroid classification without pruning

The first of our set of classification scripts repeats the process described in the previous section for each publication in our test samples. After each test sample publication has been given a query-centroid cosine similarity vector, our second classification script runs through this data, and checks whether the highest-scoring discipline corresponds to the discipline's own NOWT category. For each discipline, the classification performance was computed by dividing the total amount of correctly classified publications by the total amount of publications in that discipline in the combined test samples. The result is a per-discipline score ranging from 0 to 1 where 0 indicates no correct classifications and 1 indicates that all classifications were correct.

```
                                AGRICULTURE AND FOOD SCIENCE                                          0.451
                                                ASTRONOMY AND ASTROPHYSICS                            0.666
                                                                BASIC LIFE SCIENCES                   0.572
                                                BASIC MEDICAL SCIENCES                                0.456
                                                                BIOLOGICAL SCIENCES                   0.451
                                                                BIOMEDICAL SCIENCES                   0.297
                                CHEMISTRY AND CHEMICAL ENGINEERING                                    0.507
                                                CIVIL ENGINEERING AND CONSTRUCTION                    0.436
                                                                CLINICAL MEDICINE                     0.438
                                                COMPUTER SCIENCES                                     0.465
                                                CREATIVE ARTS, CULTURE AND MUSIC                      0.318
                                                                EARTH SCIENCES AND TECHNOLOGY         0.501
                                                ECONOMICS AND BUSINESS                                0.408
                                                EDUCATIONAL SCIENCES ELECTRICAL ENGINEERING AND TELECOMMUNICATION                    0.487                                                0.452
                                ENERGY SCIENCE AND TECHNOLOGY                                         0.343
                                                ENVIRONMENTAL SCIENCES AND TECHNOLOGY                 0.267
                                                GENERAL AND INDUSTRIAL ENGINEERING                    0.142
                                                HEALTH SCIENCES                                       0.351
                                                HISTORY, PHILOSOPHY AND RELIGION                      0.370
                                                INFORMATION AND COMMUNICATION SCIENCES                0.347
                                INSTRUMENTS AND INSTRUMENTATION                                       0.335
                                                LANGUAGE AND LINGUISTICS                              0.456
                                                                LAW AND CRIMINOLOGY                   0.349
                                                LITERATURE                                            0.534
                                                MANAGEMENT AND PLANNING                               0.264
                                                                MATHEMATICS                           0.580
                                MECHANICAL ENGINEERING AND AEROSPACE                                  0.376
                                                PHYSICS AND MATERIALS SCIENCE POLITICAL SCIENCE AND PUBLIC ADMINISTRATION                    0.427                                                0.535
                                                PSYCHOLOGY                                            0.496
                                                SOCIOLOGY AND ANTHROPOLOGY                            0.223
                                                                STATISTICAL SCIENCES                  0.435
                                                DISCIPLINE MEAN                                       0.416
                                                                PUBLICATION MEAN                      0.448
```

Figure 17: classification performance per discipline using the complete centroids.

Our first classification run of our combined test samples calculated publications' cosine similarity with the newly computed discipline centroids without pruning. The resulting classification performance table can be found in figure 17. A mean discipline classification performance and a mean publication classification performance have also been computed by averaging the classification performance of disciplines for the former and dividing the total amount of correctly classified publications by the

combined size of the test samples for the latter. The difference in these two values is due to the fact that discipline sizes are not uniform.

While a classification performance of 0.45, meaning that 45% of the publications have been classified correctly, may not sound impressive, one has to take into account that this classification was not a simple choice between two categories, but instead 33 disciplines. The chance of choosing the correct one out of 33 evenly weighed categories is roughly 3%. In this light, our classification performance is more than adequate.

The disciplinary classification performances in figure 17 do contain several peculiarities we must address. In particular, several disciplines score low compared to the discipline mean, most prominently *general and industrial engineering*, but others as well. This may indicate that these disciplines are not well-defined in terms of their noun phrase occurrence, and that publications in these discipline contain a wider range of noun phrases than those in better-performing disciplines. One could ask whether the *general engineering* category is, in fact, too general to be considered a proper discipline. *Astronomy and astrophysics*, on the other hand, seems very well-defined, which was to be expected considering its clustering behaviour in section 5.1.

7.3: Similar-centroid classification with frequency relevance threshold

While we obtained a satisfying result using the similar-centroid classification without pruning classification method, computing the cosine similarities for each test publication – discipline centroid pair was computationally expensive. As we described in section 4.5 and chapter 6, pruning the discipline centroids of low-relevance noun phrases may reduce the amount of terms used in these centroids dramatically. There may even be cause to hope that the classification performance will be improved by these pruning operations, as the discipline centroids will contain less 'noise'. While the focus of this research is not on the computational complexity of our scripts and algorithms, pruning the discipline centroids makes them easier to interpret and more representative of the cognitive content of disciplines by removing noun phrases which do not reflect unique cognitive content in the first place. We use frequency relevance because occurrence relevance has been shown to undermine our cosine similarity as shown in section 6.2.

AGRICULTURE AND FOOD SCIENCE	ASTRONOMY AND ASTROPHYSICS	BASIC LIFE SCIENCES
0.449	0.660	0.565
BASIC MEDICAL SCIENCES	BIOLOGICAL SCIENCES	BIOMEDICAL SCIENCES
0.456	0.446	0.293
CHEMISTRY AND CHEMICAL ENGINEERING	CIVIL ENGINEERING AND CONSTRUCTION	CLINICAL MEDICINE
0.492	0.434	0.439
COMPUTER SCIENCES	CREATIVE ARTS, CULTURE AND MUSIC	EARTH SCIENCES AND TECHNOLOGY
0.454	0.315	0.493
ECONOMICS AND BUSINESS	EDUCATIONAL SCIENCES	ELECTRICAL ENGINEERING AND TELECOMMUNICATION
0.404	0.485	0.445
ENERGY SCIENCE AND TECHNOLOGY	ENVIRONMENTAL SCIENCES AND TECHNOLOGY	GENERAL AND INDUSTRIAL ENGINEERING
0.342	0.263	0.148
HEALTH SCIENCES	HISTORY, PHILOSOPHY AND RELIGION	INFORMATION AND COMMUNICATION SCIENCES
0.347	0.362	0.349
INSTRUMENTS AND INSTRUMENTATION	LANGUAGE AND LINGUISTICS	LAW AND CRIMINOLOGY
0.340	0.460	0.353
LITERATURE	MANAGEMENT AND PLANNING	MATHEMATICS
0.534	0.263	0.570
MECHANICAL ENGINEERING AND AEROSPACE	PHYSICS AND MATERIALS SCIENCE	POLITICAL SCIENCE AND PUBLIC ADMINISTRATION
0.362	0.418	0.534
PSYCHOLOGY	SOCIOLOGY AND ANTHROPOLOGY	STATISTICAL SCIENCES
0.490	0.223	0.430
DISCIPLINE MEAN	PUBLICATION MEAN	
0.413	0.442	

Figure 18: classification performance per discipline with frequency relevance pruning of noun phrases in centroids, threshold = 2.

Practically, the inner workings of the methods with and without relevance pruning are identical. The only difference between the two stems from the discipline centroids used. With pruning, the discipline centroids are first pruned of noun phrases that do not score above the relevance threshold in any discipline. This pruning procedure is identical to the one described in section 6.2, again with a pruning

threshold of 2. The results of this with-pruning classification on the combined test samples can be found in figure 18.

Perhaps unsurprisingly, we find that the classification performance is similar to the without-pruning method. This was to be expected given that the pruning of the discipline centroids has been demonstrated to leave the overall structure of the discipline similarity network intact. Both the discipline mean classification performance and publication mean classification performance appear to be slightly lower than without pruning, meaning that we did not find that pruning the centroids led to an improved classification. However, reducing the amount of noun phrases used in the classification makes the landscape easier to interpret. This holds true both from a computational point of view, as using fewer noun phrases makes calculations faster, but also from a human point of view. Removing low-relevance noun phrases and presenting high-relevance noun phrases for disciplines provides context and makes our method less opaque to the less technically minded.

8: Conclusions

In this thesis, we set out to investigate whether we can find useful structure in the science space using title and abstract text of publications. For a disciplinary similarity structure to be useful, it needs to be consistent over time and reveal meaningful patterns that reflect reality. We need to be able to recognise the most relevant textual content of disciplines and to weed out irrelevant noun phrases. Finally, a structure is not useful if it cannot be put to use – it must allow us to classify publications with reasonable accuracy. Because of these conditions, we posed the following research sub-questions:

1. *Can the proposed method using title and abstract noun phrases differentiate between disciplines and find a consistent structure in a network of disciplines?*
2. *For each discipline, which are its most relevant noun phrases, and can low-relevance noun phrases be removed while preserving the overall disciplinary structure?*
3. *How accurately can disciplinary publications be classified based on the noun phrases appearing in their titles and abstracts?*

To answer these, we processed over 7 million scientific articles from the Web of Science database, divided among 33 disciplinary categories of varying size based on the NOWT subject categories (NOWT, 2010). Each document and discipline was processed as a vector in a high-dimensional *vector space model* (Salton, 1989), allowing us to establish their similarity based on their vector angles. Further, we presented two approaches to calculating the relevance of noun phrases to disciplines, to find the cognitive content of disciplines. We closed with classifying 1.4 million test publications, both with and without pruning low-relevance noun phrases.

The between-discipline similarity structure that we found is remarkably stable over time, as presented in chapter 5 and displayed in figure 6. We found consistent clusters of related disciplines representing the STEM fields, life sciences, and humanities, as well as a fourth cluster containing only *astronomy and astrophysics* which seems dissimilar to most other disciplines in its noun phrase usage. Small changes in the disciplinary similarity structure occur over the years in the sample, but correlation between discipline pair similarity scores over time remains high. We have found evidence of shifting between-discipline similarities over time (see figure 10), as well as evolving noun phrase usage within disciplines (see figure 11). This is to be expected as both science and language are evolving systems. Overall, the persistent clusters and high year-to-year correlations are evidence of a highly stable disciplinary similarity structure.

The frequency relevance measure we explored shows great promise in removing low-relevance noun phrases. Occurrence relevance favours noun phrases which are uniquely relevant for disciplines, and should be considered more a measure of specificity than of relevance in the sense that its highest scoring noun phrases are specific to single disciplines only. While this may be an effective way of finding exactly which elements set apart a discipline from the others, it is not helpful in finding those elements which link disciplines to one another and more careful selection of noun phrases is required before these results may be suitable for describing disciplines' cognitive content. Frequency relevance preserves the disciplinary similarity structures we identified earlier while allowing for a reduction of noun phrases used by three orders of magnitude. Its highest-scoring noun phrases are those that are of exceptional frequency within disciplines as compared to their frequency in the rest of the sample. Because of this, frequency relevance identifies both uniquely relevant noun phrases as well as bridging noun phrases while remaining an effective tool to weed out noun phrases whose frequency of occurrence is not exceptional in any discipline. In short, of the two measures we explored, frequency relevance is useful for pruning irrelevant terms while occurrence relevance may be promising in describing disciplines' unique content, but needs further refinement.

The classification shows that our method allows us to place publications into the disciplinary structure with reasonable accuracy. Without pruning of low-relevance noun phrases, average publication classification performance is at 44.8%, while we observed a small but consistent drop to 44.2% with pruning. Considering the high number of disciplines used in this classification, these results are very encouraging. Variations in the per-discipline classification performance do raise some concerns. Some disciplines exhibited very high classification performance, in particular *astronomy and astrophysics* (66%). Others, notably *general and industrial engineering* (15%), performed far below average. This suggests that either the language used in low-performing disciplines is less consistent, or that some of the disciplines are not as well-defined as one might have hoped. If researchers wish to further investigate the potential of using noun phrases for disciplinary classification, and expand into the realm of interdisciplinarity, the first order of business must be to define more robust disciplines and perhaps subdisciplinary structures.

Concluding, does our method allow us to *find a consistent and meaningful structure in science, and documents' place within this structure, using publications' title and abstract text?* The answer to this question is a resounding *yes*. We have shown that we can identify a consistent and stable disciplinary similarity structure, with clusters of disciplines that stay coherent over time despite evolving noun phrase usage patterns in disciplines. The clustering itself reflects well-established broader fields of science (STEM, life sciences, humanities). We can pinpoint both the specific cognitive content of each discipline as well as the overall relevant and irrelevant noun phrases using our two relevance measures, lending meaning to the disciplines and their similarities. Finally, we can effectively place publications within the disciplinary structure, even after greatly reducing the amount of noun phrases used by pruning low-relevance terms. Still, some of the disciplines we used appear to not be very well-defined. The structure should be regarded as a proof of concept, a demonstration of the validity of our method and not a goal in and of itself. Even though it reveals interesting aspects of the publications in our sample period in the Web of Science database, it is not generalizable as a be-all and end-all structure of science and derives much of its final form from the initial choice of NOWT categories as disciplines. The main and most generalizable contribution of this research is the method used to arrive at our structure. We have shown that, in principle, robust and sensible disciplinary similarity structures can be found based solely on disciplinary publications' title and abstract noun phrases. Furthermore we have shown that it is possible to place publications in this structure with encouraging accuracy, indicating that this structure is robust enough to be put to use analytically in further research.

9: Discussion

We will be the first to admit that ours is far from the first study to investigate the structure of the Web of Science database (see, for instance, Leydesdorff et al., 2013; Leydesdorff & Rafols, 2009). Likewise, the idea to use a vector space model to create networks of publications is not new (e.g. Van Raan, 2000). However, we do believe that the scope of our research makes it a unique contribution. This is mainly due to its extensive investigation of underlying between-discipline and within-discipline similarity patterns and methods for computing noun phrase relevance, as well as the classification of an extensive test sample of publications and the effects of relevance pruning thereon.

Our newfound ability to derive structure from publications' title and abstract noun phrases opens many new avenues of research. Not only can we use this method to find structure in established science systems, it may form the basis for relating other textual data, such as grey literature, grant applications, or patents, to those science systems. This will undoubtedly be challenging as the use and purpose of language is likely to differ between different types of text. Still, our method offers potential advantages over other approaches, as language is ubiquitous whereas other data (e.g. citations or funding acknowledgements) often is not. Furthermore, being able to find which disciplinary research traditions a publication builds upon, and knowing the distance between those cognitive roots, may contribute to a systematic interpretation of the interdisciplinarity, novelty value or complexity of a piece of research. While more work remains to be done before this can become a reality, we believe that our research may contribute fundamentally to such efforts. The ability to easily position new research proposals or publications in the science space and link them to relevant scientific fields and subfields should prove to be a powerful tool for government funding agencies and science policy makers as it provides handholds for selecting and steering knowledge development. The vector space model may be used to find publications with high similarity to a query, and can become a powerful discovery tool for finding related research.

A reflection on our research reveals some limitations in the structure we uncovered. While the structure remains remarkably stable over time, it is limited by the sample period as well as the sets of publications that were used to define the disciplines. This latter point is especially important, as we based our selection of disciplines on the NOWT subject categories which in turn are groupings of Web of Science subject categories. Their content depends on the indexation of journals in the database. As journals are added to the database, the cognitive content of our defined disciplines may change suddenly. We suspect that this effect is responsible for the irregular within-discipline similarity patterns of the *mechanical engineering and aerospace* discipline as displayed in figure **FIGURE6**. Another limitation of the disciplines selected is that they were constructed using only publications belonging to a single NOWT category, excluding publications in multiple categories. While it would be interesting to somehow include these multi-category publications in the structure, it is currently unclear how to best approach this – should they be assigned to multiple disciplines, possibly with some weighing to distribute their effects, or should they be treated as separate, combination categories? Finally, the choice to use the NOWT categories as disciplines could be disputed. While the resulting structure has been shown to be stable and useful, we found results which suggest that not all disciplines are equally well-defined. Selection of more robust disciplines and potentially subdisciplinary features may further improve the landscape resulting from our method as well as subsequent classification.

In purely methodological terms, there are a few issues beyond the selection of disciplines which deserve further attention. In particular, we introduced two different approaches to calculating the relevance of noun phrases per discipline. Both the occurrence relevance and frequency relevance in equations 2 and 3 depend on a parameter m, influencing the balance between absolute and relative

occurrence or frequency of noun phrases in the computation of their relevance. While this has proven to be effective, especially in the case of frequency relevance, the specific implementation of this balance renders the actual relevance scores difficult to interpret. For instance, it is unclear whether a term's low relevance is a result of its low absolute occurrence in the sample, or its low relevance occurrence, as both factors might contribute to a low relevance score. We recommend that future research improves the clarity of relevance scoring by instead using a two-step process for determining relevance. Instead, both an overall occurrence threshold and a relative occurrence threshold could be used when pruning low-relevance terms. Doing so eliminates a confusing and opaque parameter from the equations, at the cost of introducing a second but more easily interpretable pruning threshold.

The method may be further expanded by further leveraging the vector space model. As discussed earlier, not all the disciplines we used seemed equally well-defined. It may be possible to find a metric for a discipline's quality of definition. For instance, one may use the average multi-dimensional distance between the publications within a discipline and that discipline's centroid, to see how 'large' the space covered by the discipline is. Alternatively one could simply compute the average cosine similarity of a discipline's publications and its centroid as a measure of dispersion.

9.1: Further research

During the course of this research we have encountered several interesting phenomena and possible avenues for further research. While we have already touched upon the need for more robust disciplines if research into this method for structuring the science space is to continue, a second and exciting possibility is to construct a disciplinary structure in science from the ground up, similar to what Waltman & Van Eck (2012) do with direct citation data, but with noun phrase occurrence patterns instead.

If, by whatever way, we can arrive at a more robust, and more encompassing, disciplinary foundation for this type of research, we may be able to expand it beyond disciplinary classification into the realm of disciplinary integration and interdisciplinarity. At the start of this thesis project we set out to do exactly that, with the ambition to construct a method for assigning interdisciplinarity scores to publications based on their noun phrase usage, but on our way there we found a myriad of more fundamental questions that needed to be addressed first, resulting in the thesis as it lies before you now. While our theoretical framework allowed us to construct robust methods for finding between-discipline and within-discipline similarity structures and come to an adequate classification of publications, it remains largely untested as far as its notions on disciplinary integration go.

An interesting possibility for further research came up in discussion with my colleagues after the presentation of the preliminary results of this thesis. Moving away from notions of disciplinarity and disciplinary integration, one may consider using noun phrases and their occurrence in publications as a bridge between science and technology. If one is able to identify terms associated with a certain technology or instrument, one may find publications influencing or influenced by this technology by charting which publications use these terms. An example is the Hubble space telescope, a high-relevance noun phrase in the *astronomy and astrophysics* discipline. It would be trivially easy to collect all publications containing this noun phrase in their abstracts, to see which publications have benefited from the data produced by the telescope. A more complex but perhaps also more interesting way of using noun phrases to link science and innovation would be to search for precursor publications to technological breakthroughs, or to map the co-development of certain instruments and branches of experimental science.

10: References

Altan, O., Ozkan, S., Altan, A., Akbas, Y., Ayhan, V., & Ozkan, K. (2000). Effects of short-term fasting and midnight lighting on egg production traits of laying hens during summer season. *ARCHIV FUR GEFLUGELKUNDE*, *64*(2), 85–89. Retrieved from http://apps.webofknowledge.com.proxy.library.uu.nl/full_record.do?product=WOS&search_mode=AdvancedSearch&qid=5&SID=W2jdORKT8KcV4anq3oO&page=1&doc=1

Alvargonzález, D. (2011). Multidisciplinarity, Interdisciplinarity, Transdisciplinarity, and the Sciences. *International Studies in the Philosophy of Science*, *25*(4), 387–403. doi:10.1080/02698595.2011.623366

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759–1767. doi:10.1002/asi.22896

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects. *Journal of the American Society for Information Science*, *42*(1988), 252–266. doi:10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I

Bruin, R. E., & Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics*, *26*(1), 65–80. doi:10.1007/BF02016793

Choi, B. C. K., & Pak, A. W. P. (2006). Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine*, *29*(6), 351–364. doi:10.1016/j.jaac.2010.08.010

Costas, R., Zahedi, Z., & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *arXiv Preprint arXiv:1401.4321*, 30. doi:10.1002/asi.23309

Darden, L., & Maull, N. (1977). Interfield Theories. *Philosophy of Science*, *44*(1), 43–64.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, *66*(7), 1374–1387. doi:10.1002/asi.23271

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. SAGE Publications. Retrieved from https://books.google.com/books?hl=en&lr=&id=WQRwikYjh_0C&pgis=1

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. doi:10.1023/A:1022378804087

Heimeriks, G. (2013). Interdisciplinarity in biotechnology, genomics and nanotechnology. *Science and Public Policy*, *40*(1), 97–112. doi:10.1093/scipol/scs070

Hessels, L. K., & van Lente, H. (2008). Re-thinking new knowledge production: A literature review and a research agenda. *Research Policy*, *37*(4), 740–760. doi:10.1016/j.respol.2008.01.008

Hicks, D. M., & Katz, S. (1996). Where Is Science Going? *Science, Technology, & Human Values*, *21*(4), 379–406. doi:10.1177/016224399602100401

Jurka, T., & Collingwood, L. (2013). RTextTools: A Supervised Learning Package for Text Classification. *R Journal*, *5*(1), 6–12. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=20734859&AN=90616103&h=50cyR6MarAY9WE/sMi2d4KY2rxKU7dzzMDzuuThqpFgz4Zxe3x3+d7WcPDSIcW+g0t7yaqx3AXpG8xNmnd35Jg==&crl=c

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. *Philosophical Review* (Vol. II). doi:10.1119/1.1969660

Kurz, H. D. (2012). Schumpeter's new combinations: Revisiting his Theorie der wirtschaftlichen Entwicklung on the occasion of its centenary. *Journal of Evolutionary Economics*, *22*, 871–899. doi:10.1007/s00191-012-0295-z

Lélé, S., & Norgaard, R. B. (2005). Practicing Interdisciplinarity. *BioScience*, *55*(11), 967. doi:10.1641/0006-3568(2005)055[0967:PI]2.0.CO;2

Leydesdorff, L. (1997). Why words and co-words cannot map the development of the sciences. *Journal of the American Society for Information Science*, *48*(5), 418–427. doi:10.1002/(SICI)1097-4571(199705)48:5<418::AID-ASI4>3.0.CO;2-Y

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, *59*(1), 77–85. doi:10.1002/asi.20732

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589–593. doi:10.1007/s11192-012-0784-8

Leydesdorff, L., & Meyer, M. (2006). Triple Helix indicators of knowledge-based innovation systems. *Research Policy*, *35*(10), 1441–1449. doi:10.1016/j.respol.2006.09.016

Leydesdorff, L., & Rafols, I. (2009). A Global Map of Science Based on the ISI Subject. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. doi:10.1002/asi

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S &T systems*. *Book*. doi:10.1007/1-4020-2755-9_20

National Science Board. (2014). Science & Engineering Indicators. *National Science Board*, 5–7. Retrieved from http://www.nsf.gov/statistics/seind14/index.cfm/etc/pdf.htm

NOWT. (2010). Science and Technology Indicators 2010. *Netherlands Observatory of Science and Technology*.

Peters, H. P. F., & van Raan, A. F. J. (1993a). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy*, *22*(1), 23–45. doi:10.1016/0048-7333(93)90031-C

Peters, H. P. F., & van Raan, A. F. J. (1993b). Co-word-based science maps of chemical engineering. Part II: Representations by combined clustering and multidimensional scaling. *Research Policy*, *22*(1), 47–71. doi:10.1016/0048-7333(93)90032-D

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745. doi:10.1007/s11192-008-2197-2

Porter, A. L., Roessner, J. D., Cohen, A. S., & Perrault, M. (2006). Interdisciplinary research: meaning, metrics and nurture. *Research Evaluation*, *15*(3), 187–195. Retrieved from http://rev.oxfordjournals.org/content/15/3/187.short

Rinia, E. J. (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer*.

Rosenfield, P. L. (1992). The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science and Medicine*, *35*(Ii), 1343–1357. doi:10.1016/0277-9536(92)90038-R

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc. Retrieved from http://books.google.com/books?id=wb8SAQAAMAAJ&pgis=1

Schmidt, J. C. (2008). Towards a philosophy of interdisciplinarity. An attempt to provide a classification and clarification. *Poiesis & Praxis*, *5*, 53–69. doi:10.1007/s10202-007-0037-8

Schmidt, J. C. (2011). What is a problem? On problem-oriented interdisciplinarity. *Poiesis & Praxis*, *7*, 249–274. doi:10.1007/s10202-011-0091-0

Schumpeter, J. A. (1934). *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle*. Transaction Publishers. Retrieved from https://books.google.com/books?id=-OZwWcOGeOwC&pgis=1

Stehr, N., & Weingart, P. (2000). *Practising interdisciplinarity*. (N. Stehr & P. Weingart, Eds.). University of Toronto Press. Retrieved from https://books.google.nl/books?hl=en&lr=&id=MS92DLz4gksC&oi=fnd&pg=PR11&ots=XQ2lcitref&sig=bFM8_pHoSHJ7508qOf_ts32cYFI

Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L., … Trochim, W. (2003). Evaluating transdisciplinary science. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco*, *5 Suppl 1*(December), S21–S39. doi:10.1080/14622200310001625555

Turner, S. (2000). What are disciplines? And how is interdisciplinarity different. In N. Stehr & P. Weingart (Eds.), *Practising interdisciplinarity* (pp. 46–65). University of Toronto Press.

Van den Besselaar, P., & Heimeriks, G. (2001). Disciplinary, Multidisciplinary, Interdisciplinary - Concepts and Indicators. In *Scientometrics and Informetrics* (pp. 1–9).

Van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377–393.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. doi:10.1007/s11192-009-0146-3

Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer, 1–5. Retrieved from http://arxiv.org/abs/1109.2058

Van Eck, N. J., Waltman, L., Noyons, E. C. M., & Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, *82*(3), 581–596. doi:10.1007/s11192-010-0173-0

Van Raan, A. F. J. (2000). The Interdisciplinary Nature of Science: Theoritcal Framework and Bibliometric-Empirical Approach. In N. Stehr & P. Weingart (Eds.), *Practising interdisciplinarity* (pp. 66–78). University of Toronto Press.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., … Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14–26. doi:10.1016/j.joi.2010.06.004

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. doi:10.1002/asi.22748

Weingart, P. (2000). Interdisciplinarity: The paradoxical discourse. In N. Stehr & P. Weingart (Eds.), *Practising interdisciplinarity* (pp. 25–41). University of Toronto Press.

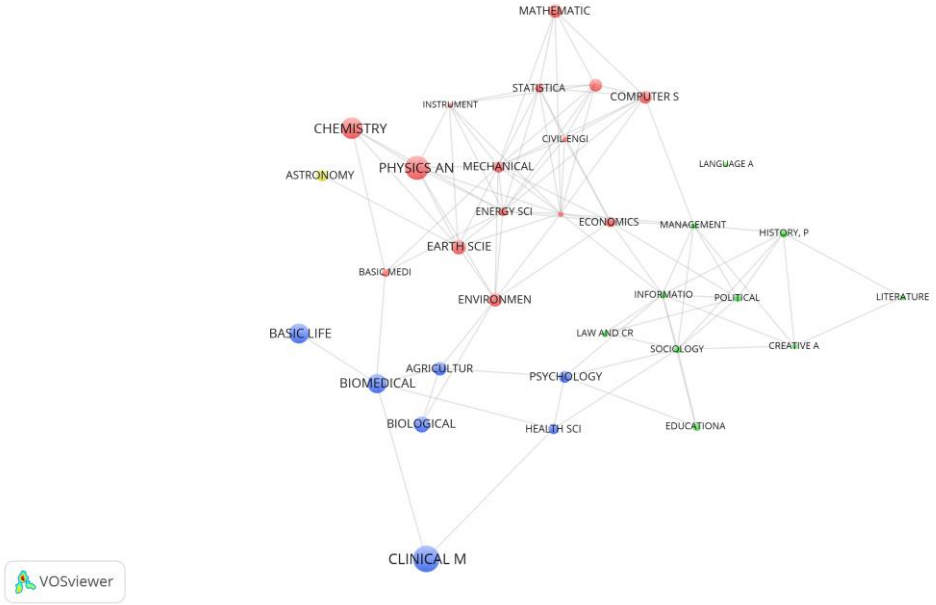
Appendix A: Complete sample information

This table lists, for each discipline, the number of publications in that discipline for each year in the sample, as well as year and discipline totals.

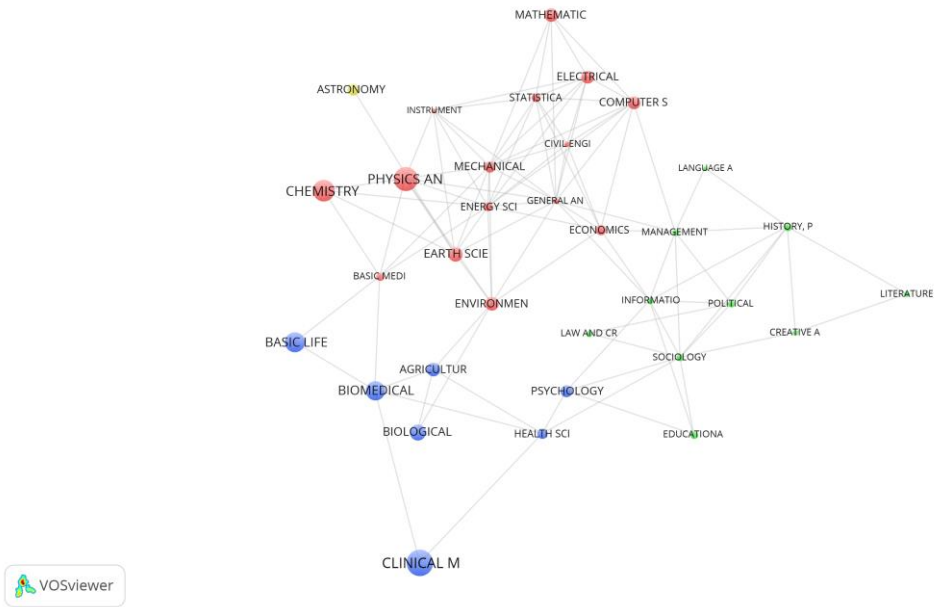
Discipline name	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
Agriculture and food science	12781	12424	13188	14244	14750	14616	15835	18866	21686	22549	23398	184337
Astronomy and astrophysics	8044	9036	8511	9317	9531	9140	9581	9643	9652	10208	10817	103480
Basic life sciences	48369	48181	48349	49392	51089	51361	52558	54786	54514	54514	57127	570240
Basic medical sciences	2243	2328	2693	2734	3160	3621	3932	4185	5140	6286	7238	43560
Biological sciences	21571	21743	23000	23287	24681	25823	27578	29793	32253	31881	33900	295510
Biomedical sciences	40613	39877	39439	40544	41999	43292	44946	47261	49920	50106	52658	490655
Chemistry and chemical engineering	60964	62773	64036	66922	71742	74377	76289	78456	80747	82388	85998	804692
Civil engineering and construction	960	941	938	1066	1191	1269	1427	1979	2160	2378	2534	16843
Clinical medicine	114482	115520	117685	122550	126798	134386	140905	154381	167520	175767	183499	1553493
Computer sciences	11748	9957	14036	21341	26335	30719	28365	9914	10886	11145	11653	186099
Creative arts, culture and music	876	976	1127	1158	1189	1586	1817	2209	2887	3273	3589	20687
Earth sciences and technology	17592	18295	18558	19788	20752	21163	22282	23595	25226	26430	28241	241922
Economics and business	6236	6045	6080	6277	6452	6880	7358	8799	11243	11586	12185	89141
Educational sciences	2244	2064	2096	2142	2374	2456	2768	3599	4653	5511	6111	36018
Electrical engineering and telecommunication	11412	10572	10824	12046	12512	13948	14361	16501	17320	18369	19997	157862
Energy science and technology	2968	3151	3048	3221	2961	3479	3147	3530	4156	4230	4239	38130
Environmental sciences and technology	12472	13826	14247	15885	16515	17984	19743	21261	24571	25817	27066	209387
General and industrial engineering	635	678	758	862	888	918	897	1469	1710	1725	2124	12664
Health sciences	6326	6526	6718	6895	7390	8084	9009	10550	12303	13365	14725	101891
History, philosophy and religion	2480	2690	2933	3205	3223	3857	4386	5212	6955	7402	8118	50461
Information and communication sciences	1025	995	1081	1127	1192	1413	1456	1735	2081	2305	2409	16819
Instruments and instrumentation	538	542	502	597	626	504	603	601	871	865	987	7236
Language and linguistics	518	553	634	663	793	852	946	1257	1797	1941	2167	12121
Law and criminology	1259	1344	1202	1209	1312	1380	1472	1771	2398	2690	3259	19296
Literature	672	702	770	802	931	1161	1268	1537	1692	1757	1774	13066
Management and planning	1062	1188	1175	1299	1421	1540	1599	1868	2496	3178	3556	20382
Mathematics	14052	14775	15760	17521	17760	19572	21702	23918	26292	27967	28312	227631
Mechanical engineering and aerospace	7750	8343	8267	8477	8859	9130	10300	11028	11694	12131	12181	108160
Physics and materials science	86309	85648	91435	93725	102356	108606	115794	115707	119007	118940	118635	1156162
Political science and public administration	1688	1686	1748	1890	2067	2208	2412	2758	3535	4043	4336	28371
Psychology	9848	9676	9718	10083	10326	10821	11823	12993	14183	14874	15607	129952
Sociology and anthropology	2017	2016	2051	2113	2137	2324	2514	2976	3820	4005	4482	30455
Statistical sciences	3168	3064	3068	3138	3413	3533	3916	4547	5046	5350	5212	43455
Total	514922	518135	535675	565520	598725	632003	662989	688685	740414	764976	798134	7020178

Appendix B: Per-year discipline similarity maps

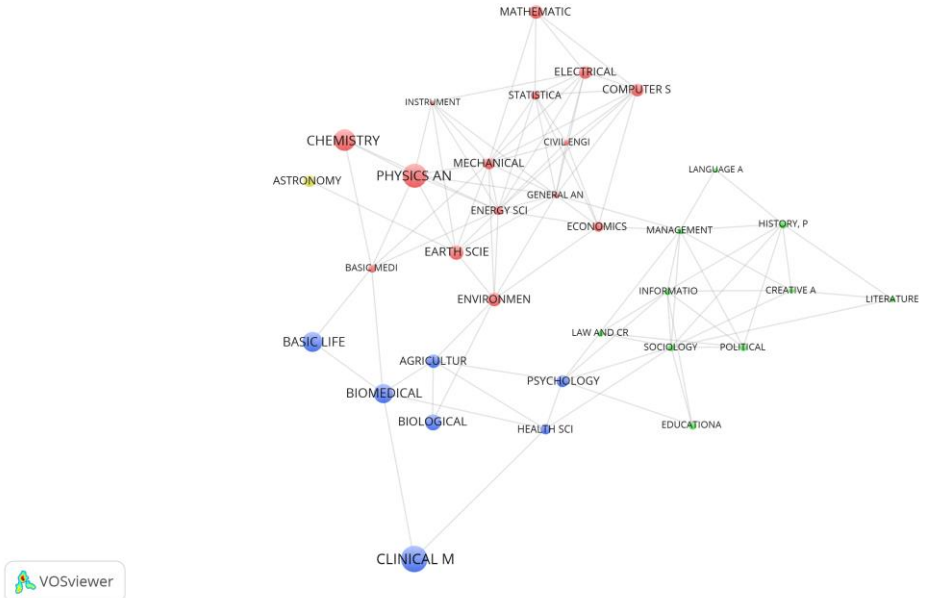
This appendix contains all per-year VOSViewer maps of the structure uncovered in chapter 5, as well as a map of the complete sample disciplinary similarity structure.



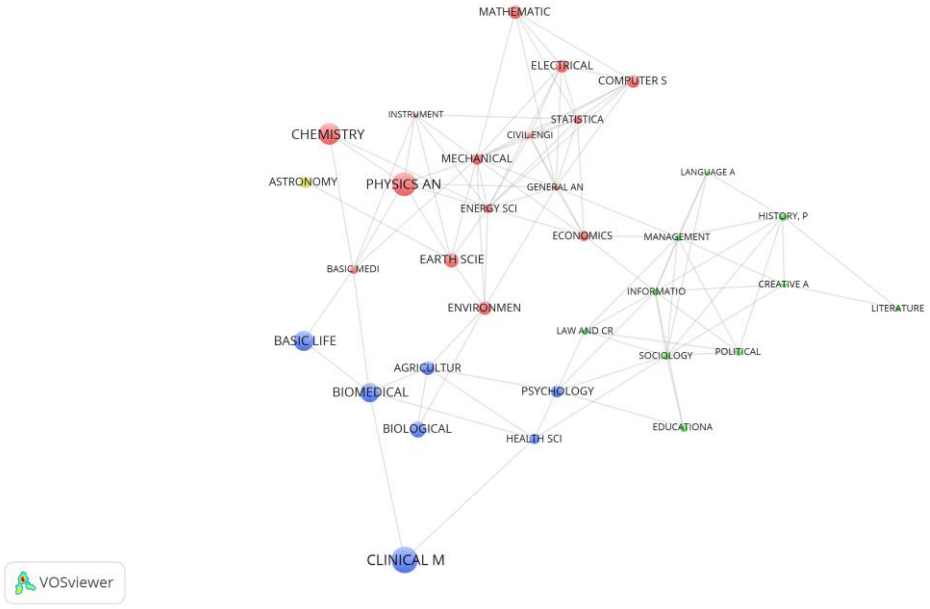
2000



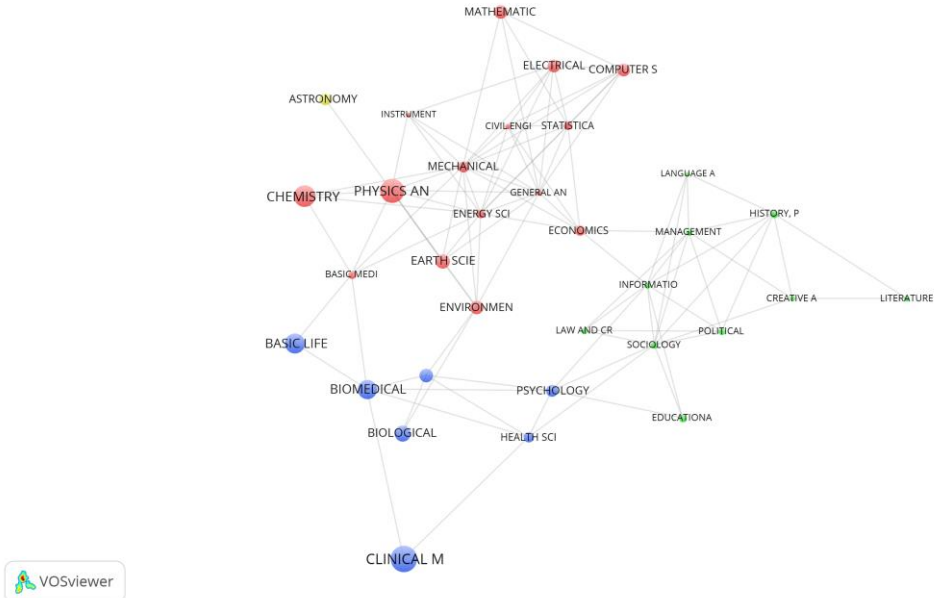
2001


MATHEMATIC
ELECTRICAL
COMPUTER S
STATISTICA
INSTRUMENT
CHEMISTRY
CIVILENGI
MECHANICAL
ASTRONOMY
PHYSICS AN
LANGUAGE A
GENERAL AN
ENERGY SCI
ECONOMICS
MANAGEMENT
HISTORY, P
EARTH SCIE
BASICMEDI
INFORMATIO
CREATIVE A
LITERATURE
ENVIRONMEN
LAW AND CR
BASIC LIFE
SOCIOLOGY
POLITICAL
AGRICULTUR
PSYCHOLOGY
BIOMEDICAL
BIOLOGICAL
HEALTH SCI
EDUCATIONA
CLINICAL M
VOSviewer

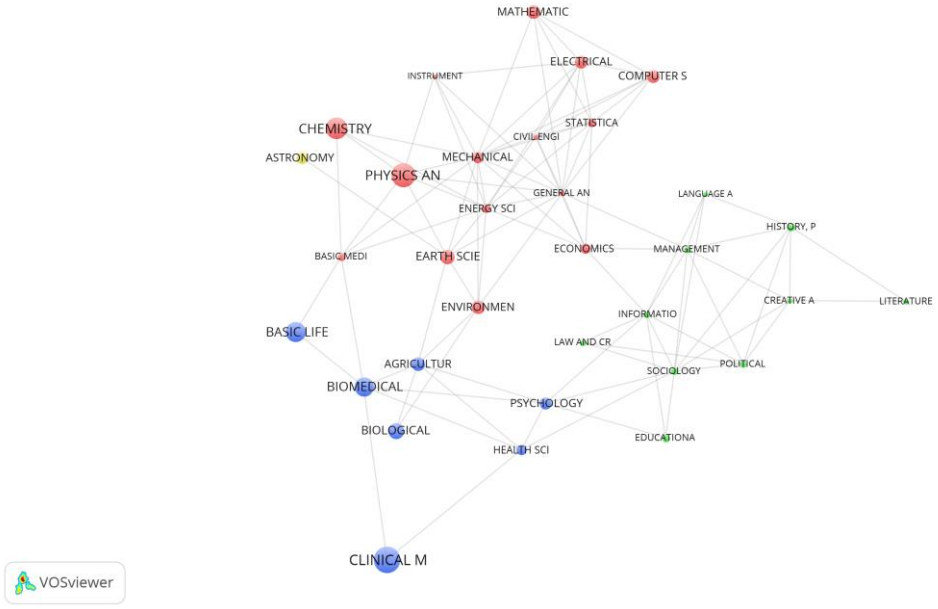
2002


MATHEMATIC
ELECTRICAL
COMPUTER S
INSTRUMENT
STATISTICA
CHEMISTRY
CIVILENGI
MECHANICAL
ASTRONOMY
PHYSICS AN
LANGUAGE A
GENERAL AN
ENERGY SCI
HISTORY, P
ECONOMICS
MANAGEMENT
EARTH SCIE
BASICMEDI
CREATIVE A
INFORMATIO
LITERATURE
ENVIRONMEN
LAW AND CR
BASIC LIFE
SOCIOLOGY
POLITICAL
AGRICULTUR
BIOMEDICAL
PSYCHOLOGY
BIOLOGICAL
EDUCATIONA
HEALTH SCI
CLINICAL M
VOSviewer

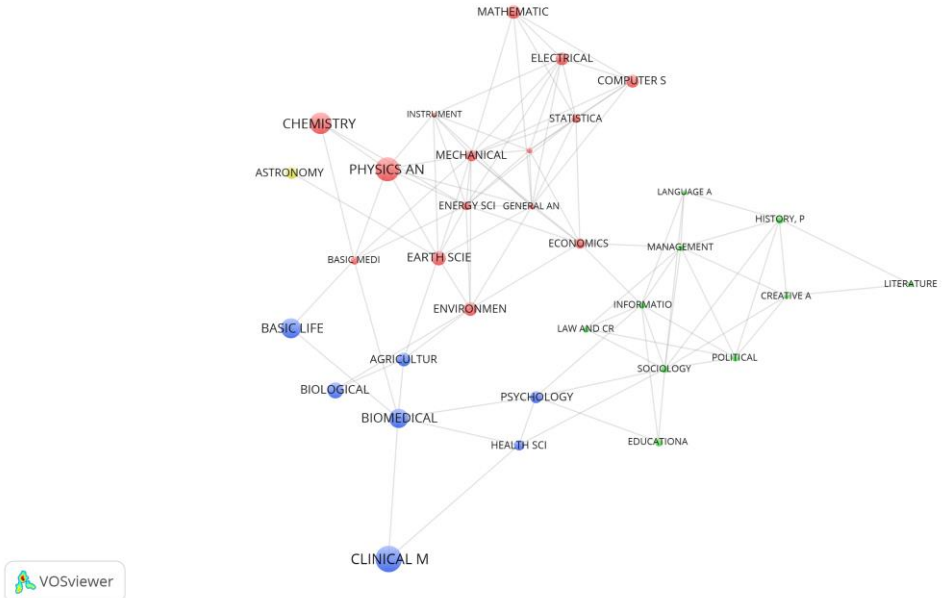
2003


MATHEMATIC
ELECTRICAL
COMPUTER S
ASTRONOMY
INSTRUMENT
CIVILENGI
STATISTICA
MECHANICAL
CHEMISTRY
PHYSICS AN
GENERAL AN
LANGUAGE A
ENERGY SCI
HISTORY, P
ECONOMICS
MANAGEMENT
EARTH SCIE
BASICMEDI
INFORMATIO
CREATIVE A
LITERATURE
ENVIRONMEN
LAW AND CR
POLITICAL
BASIC LIFE
SOCIOLOGY
BIOMEDICAL
PSYCHOLOGY
EDUCATIONA
BIOLOGICAL
HEALTH SCI
CLINICAL M
VOSviewer

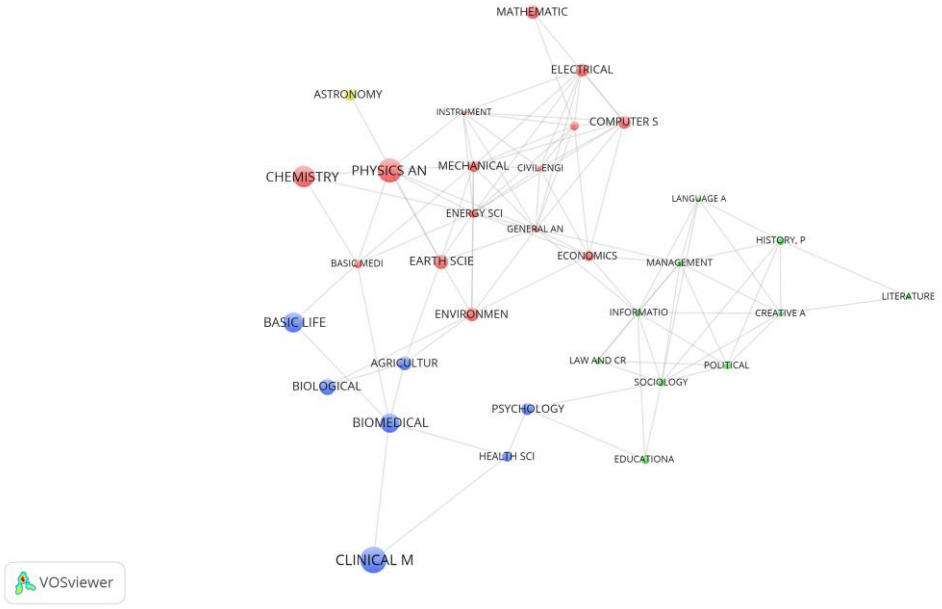
2004


MATHEMATIC
ELECTRICAL
INSTRUMENT
COMPUTER S
CHEMISTRY
STATISTICA
CIVILENGI
ASTRONOMY
MECHANICAL
PHYSICS AN
GENERAL AN
LANGUAGE A
ENERGY SCI
HISTORY, P
BASICMEDI
EARTH SCIE
ECONOMICS
MANAGEMENT
CREATIVE A
LITERATURE
ENVIRONMEN
INFORMATIO
BASIC LIFE
LAW AND CR
AGRICULTUR
SOCIOLOGY
POLITICAL
BIOMEDICAL
PSYCHOLOGY
BIOLOGICAL
EDUCATIONA
HEALTH SCI
CLINICAL M
VOSviewer

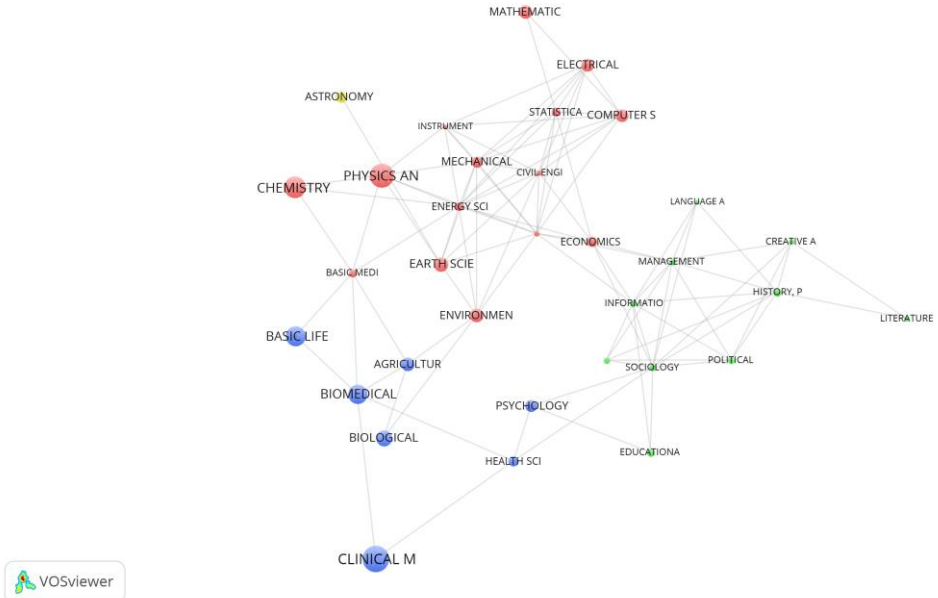
2005


MATHEMATIC
ELECTRICAL
COMPUTER S
CHEMISTRY
INSTRUMENT
STATISTICA
MECHANICAL
ASTRONOMY
PHYSICS AN
LANGUAGE A
HISTORY, P
ENERGY SCI
GENERAL AN
ECONOMICS
MANAGEMENT
BASIC MEDI
EARTH SCIE
LITERATURE
CREATIVE A
ENVIRONMEN
INFORMATIO
BASIC LIFE
LAW AND CR
AGRICULTUR
POLITICAL
SOCIOLOGY
BIOLOGICAL
PSYCHOLOGY
BIOMEDICAL
HEALTH SCI
EDUCATIONA
CLINICAL M
VOSviewer

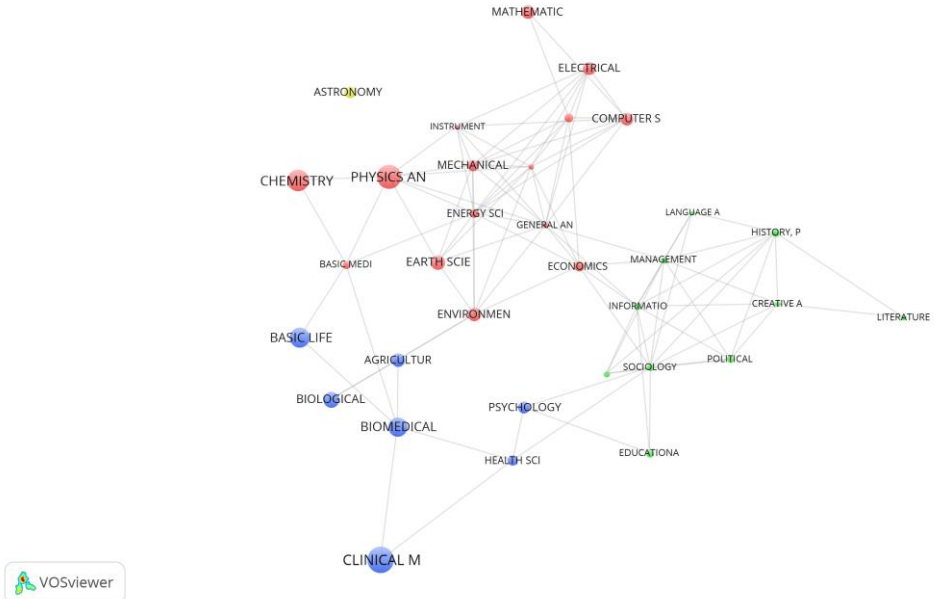
2006


MATHEMATIC
ELECTRICAL
ASTRONOMY
INSTRUMENT
COMPUTER S
CHEMISTRY
PHYSICS AN
MECHANICAL
CIVIL ENGI
LANGUAGE A
ENERGY SCI
GENERAL AN
HISTORY, P
BASIC MEDI
EARTH SCIE
ECONOMICS
MANAGEMENT
LITERATURE
ENVIRONMEN
INFORMATIO
CREATIVE A
BASIC LIFE
AGRICULTUR
LAW AND CR
POLITICAL
SOCIOLOGY
BIOLOGICAL
PSYCHOLOGY
BIOMEDICAL
HEALTH SCI
EDUCATIONA
CLINICAL M
VOSviewer

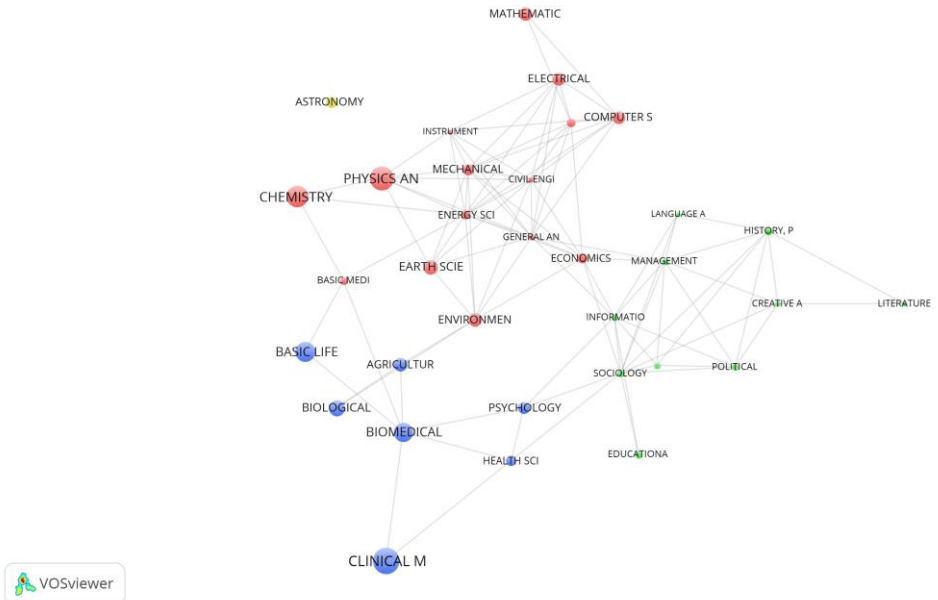
2007


MATHEMATIC
ELECTRICAL
ASTRONOMY
STATISTICA
COMPUTER S
INSTRUMENT
MECHANICAL
CIVIL ENGI
PHYSICS AN
CHEMISTRY
ENERGY SCI
LANGUAGE A
ECONOMICS
CREATIVE A
MANAGEMENT
EARTH SCIE
BASIC MEDI
HISTORY, P
INFORMATIO
ENVIRONMEN
LITERATURE
BASIC LIFE
AGRICULTUR
SOCIOLOGY
POLITICAL
BIOMEDICAL
PSYCHOLOGY
BIOLOGICAL
EDUCATIONA
HEALTH SCI
CLINICAL M
VOSviewer

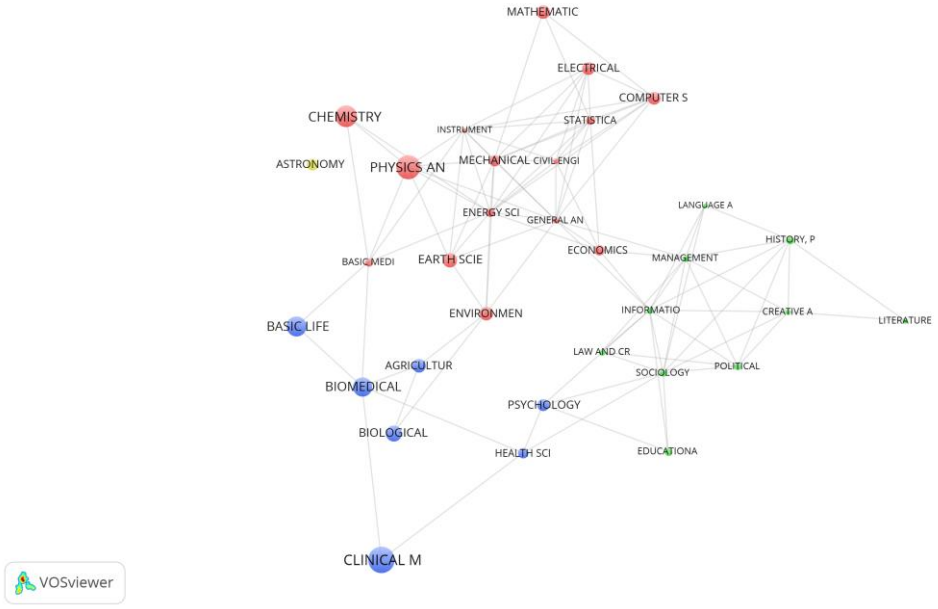
2008


MATHEMATIC
ELECTRICAL
ASTRONOMY
COMPUTER S
INSTRUMENT
MECHANICAL
CHEMISTRY
PHYSICS AN
ENERGY SCI
GENERAL AN
LANGUAGE A
HISTORY, P
BASIC MEDI
EARTH SCIE
ECONOMICS
MANAGEMENT
INFORMATIO
CREATIVE A
LITERATURE
ENVIRONMEN
BASIC LIFE
AGRICULTUR
SOCIOLOGY
POLITICAL
BIOLOGICAL
PSYCHOLOGY
BIOMEDICAL
EDUCATIONA
HEALTH SCI
CLINICAL M
VOSviewer

2009



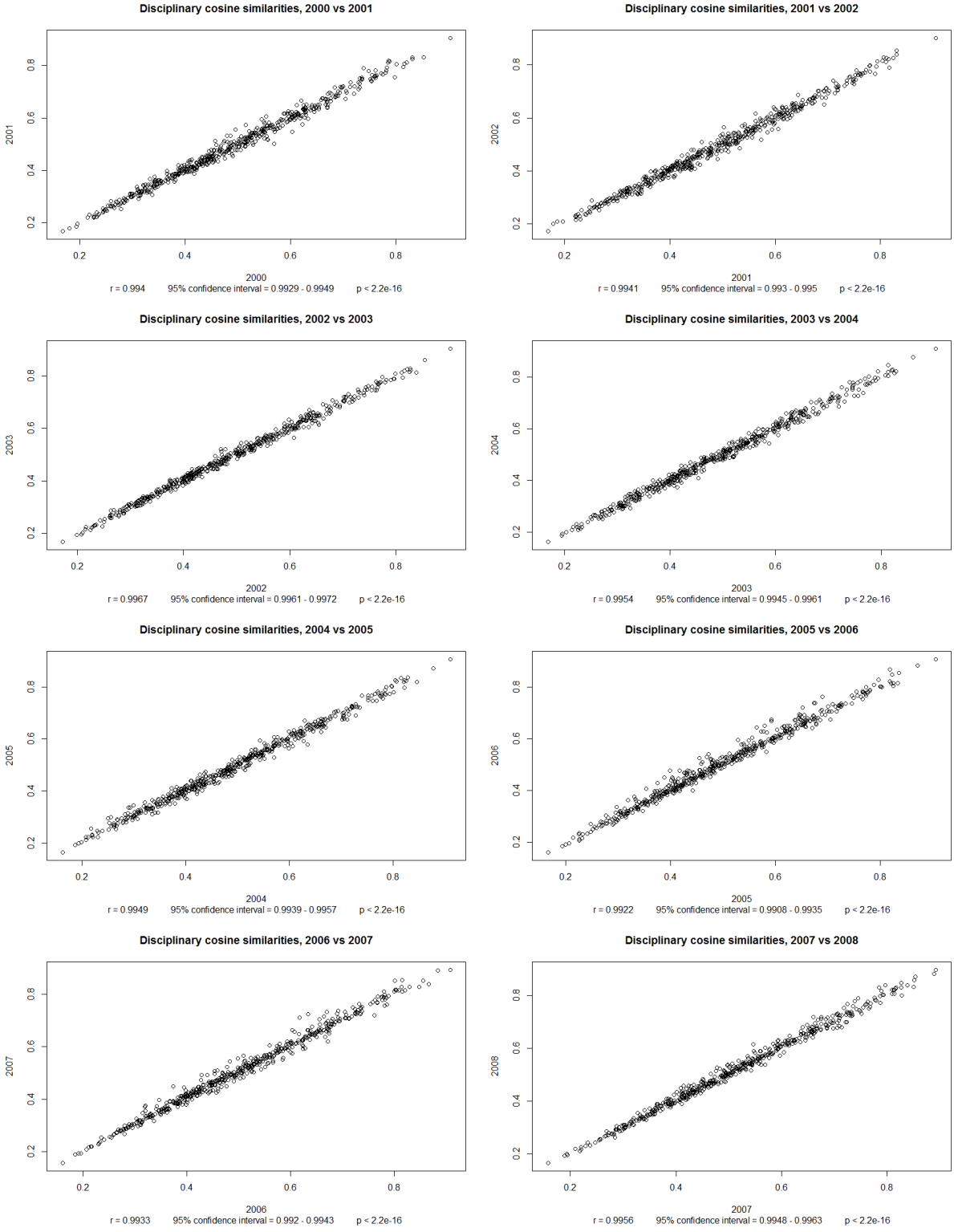
2010

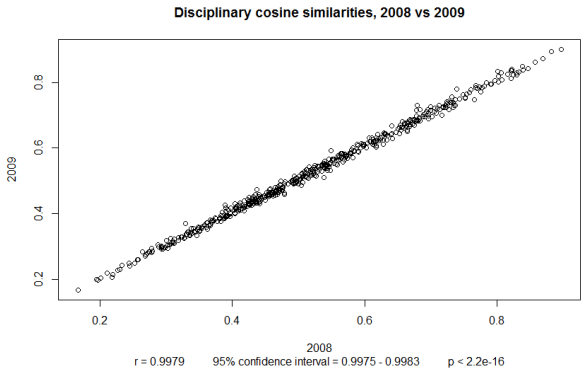


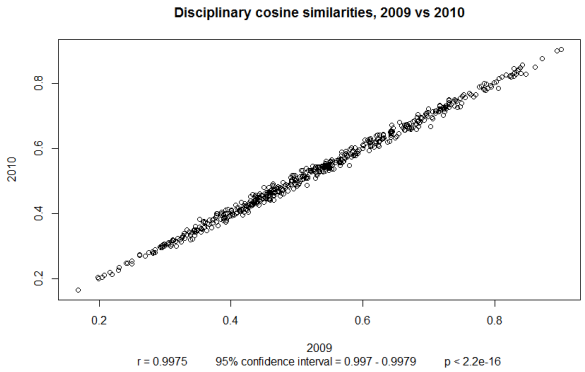
Complete sample

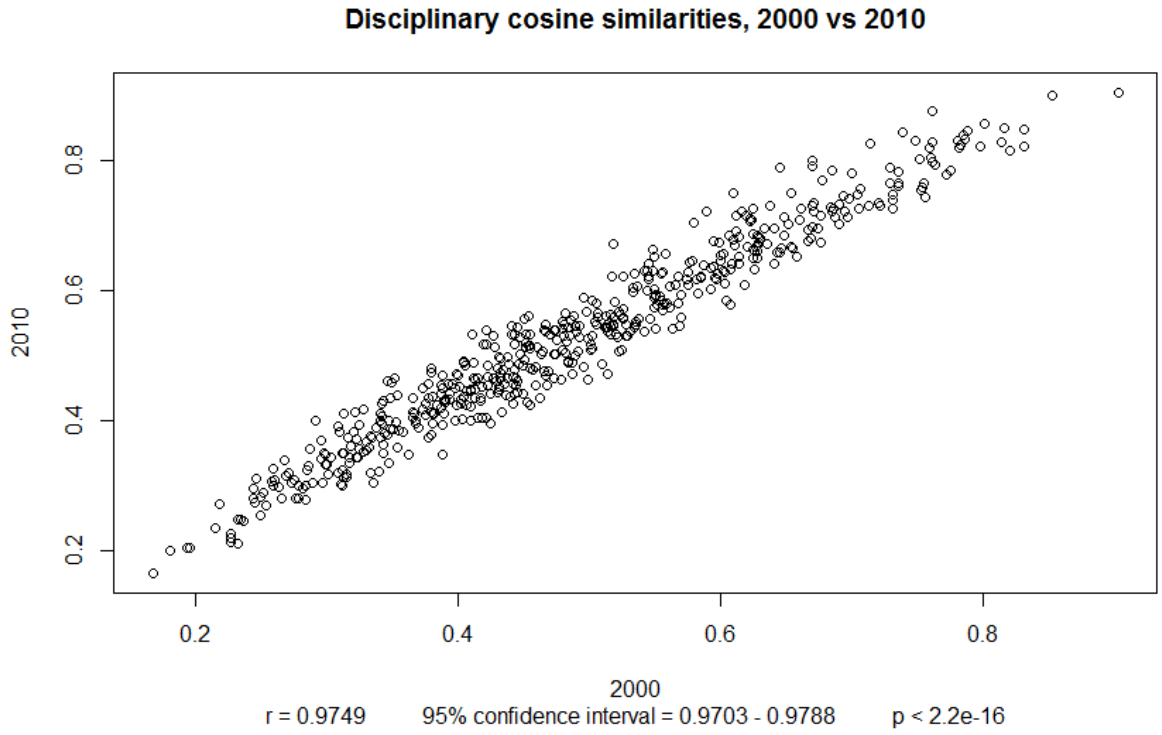
Appendix C: Correlation test results

This appendix contains all subsequent-year correlation tests and graphs made for chapter 5.2, as well as a plot comparing the disciplinary similarity data of 2000 with that of 2010.



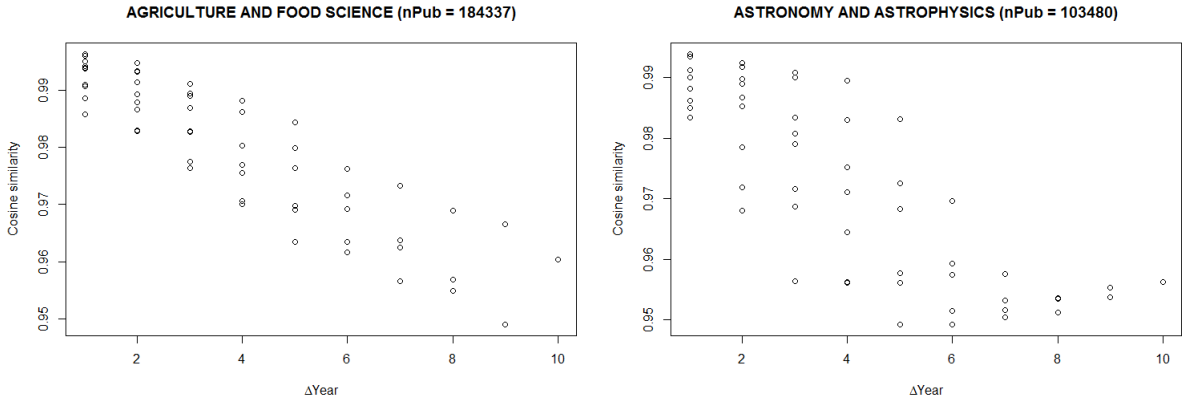

Disciplinary cosine similarities, 2008 vs 2009
2009
0.2
0.4
0.6
0.8
2008
r = 0.9979 95% confidence interval = 0.9975 - 0.9983 p < 2.2e-16

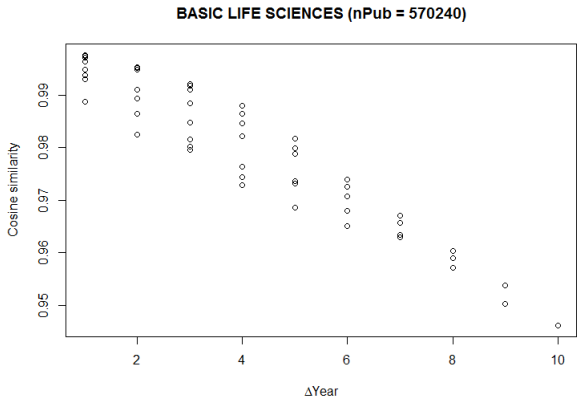

Disciplinary cosine similarities, 2009 vs 2010
2010
0.2
0.4
0.6
0.8
2009
r = 0.9975 95% confidence interval = 0.997 - 0.9979 p < 2.2e-16

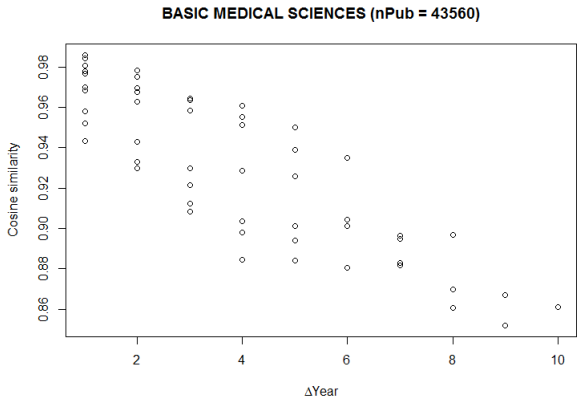

Disciplinary cosine similarities, 2000 vs 2010
2010
0.2
0.4
0.6
0.8
2000
r = 0.9749 95% confidence interval = 0.9703 - 0.9788 p < 2.2e-16

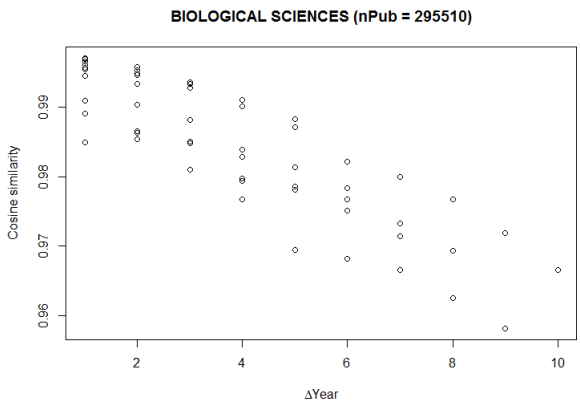
Appendix D: Within-discipline similarity plots

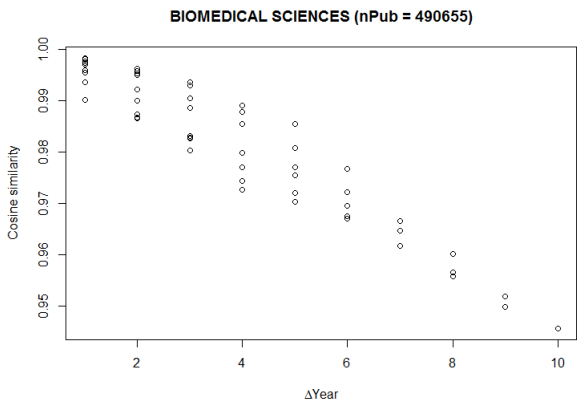
This appendix contains all within-discipline similarity plots made for chapter 5.3.

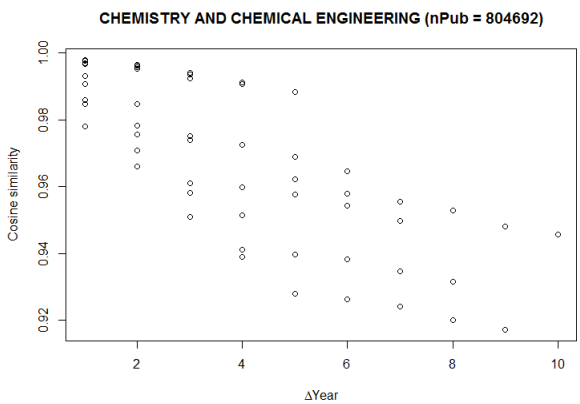


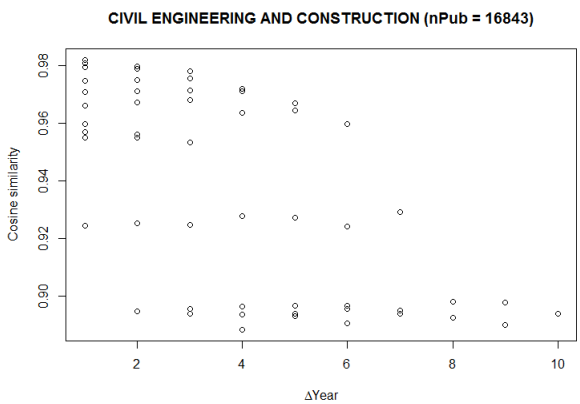


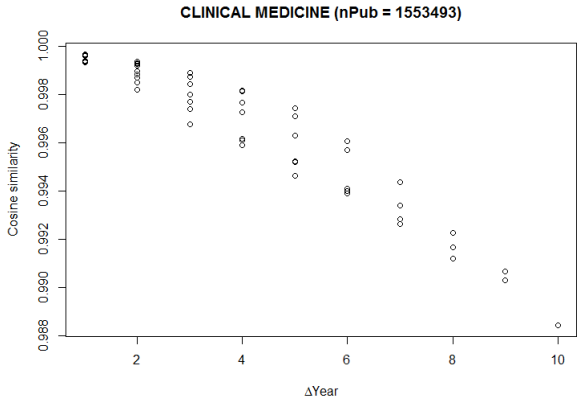


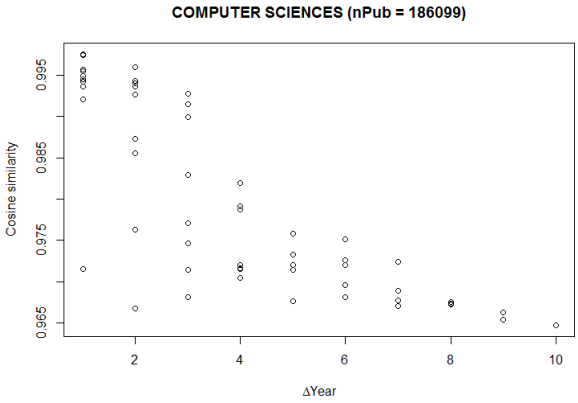


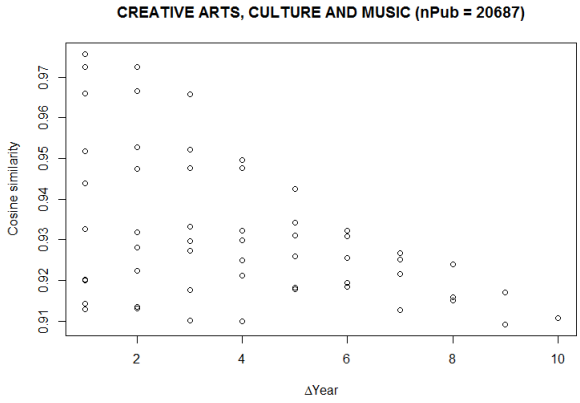


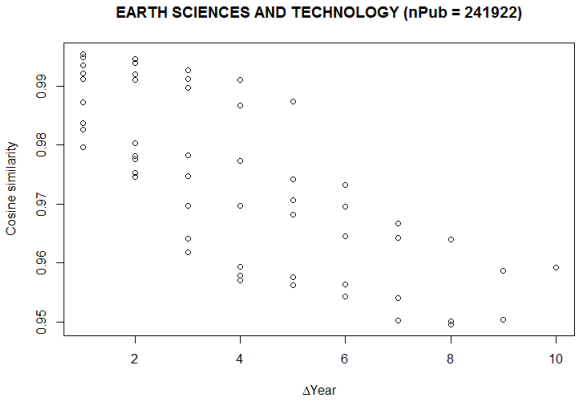


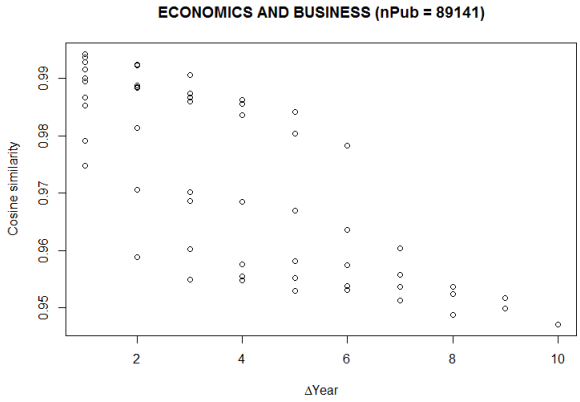


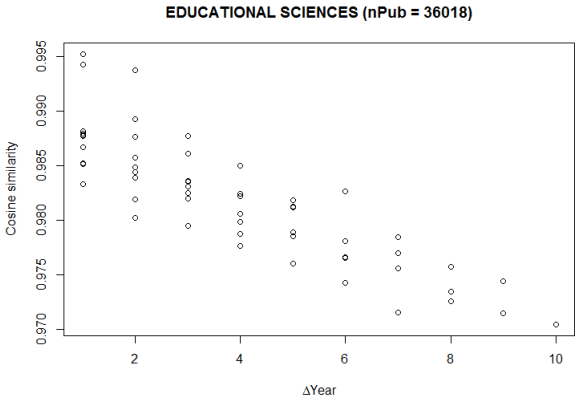

CLINICAL MEDICINE (nPub = 1553493)
Cosine similarity
ΔYear

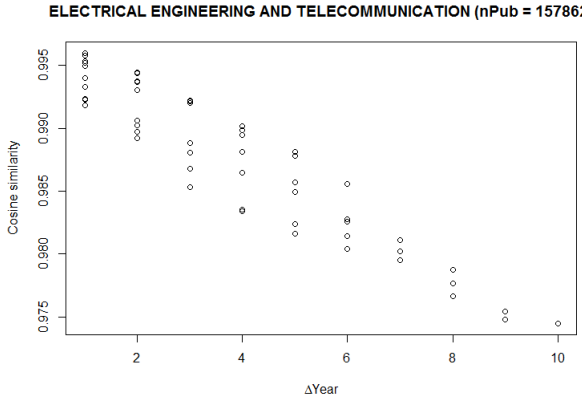

COMPUTER SCIENCES (nPub = 186099)
Cosine similarity
ΔYear

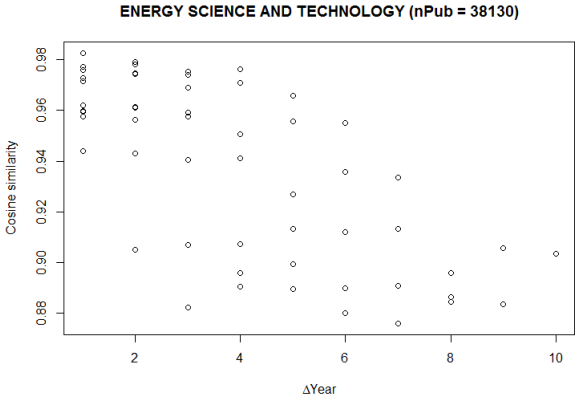

CREATIVE ARTS, CULTURE AND MUSIC (nPub = 20687)
Cosine similarity
ΔYear

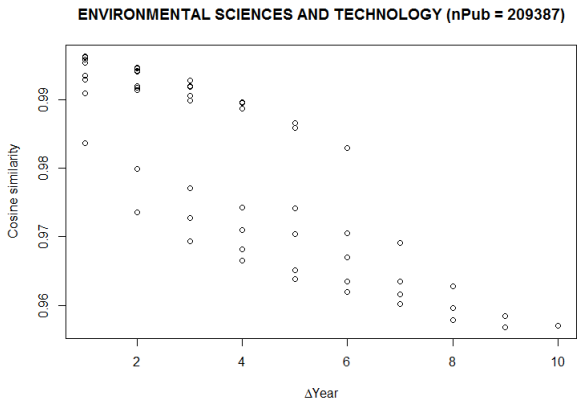

EARTH SCIENCES AND TECHNOLOGY (nPub = 241922)
Cosine similarity
ΔYear

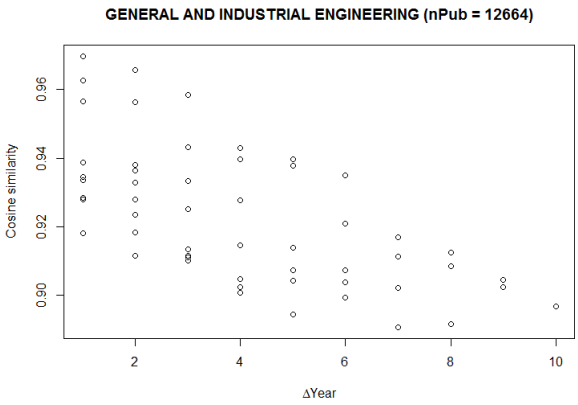

ECONOMICS AND BUSINESS (nPub = 89141)
Cosine similarity
ΔYear

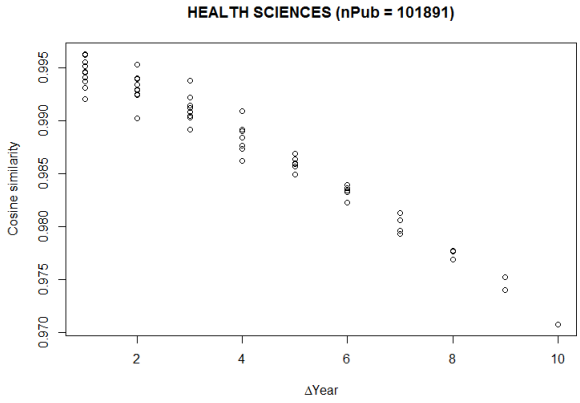

EDUCATIONAL SCIENCES (nPub = 36018)
Cosine similarity
ΔYear

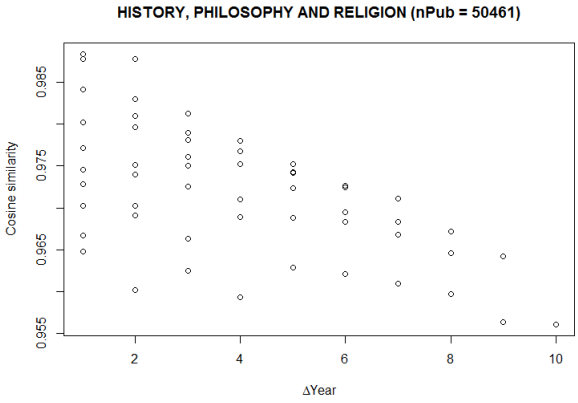

ELECTRICAL ENGINEERING AND TELECOMMUNICATION (nPub = 157862)
Cosine similarity
ΔYear

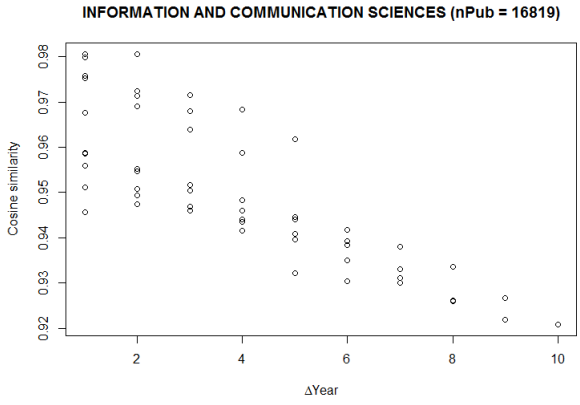

ENERGY SCIENCE AND TECHNOLOGY (nPub = 38130)
Cosine similarity
ΔYear

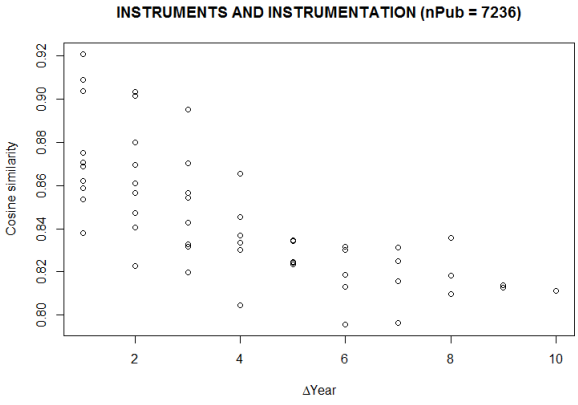

ENVIRONMENTAL SCIENCES AND TECHNOLOGY (nPub = 209387)
Cosine similarity
ΔYear

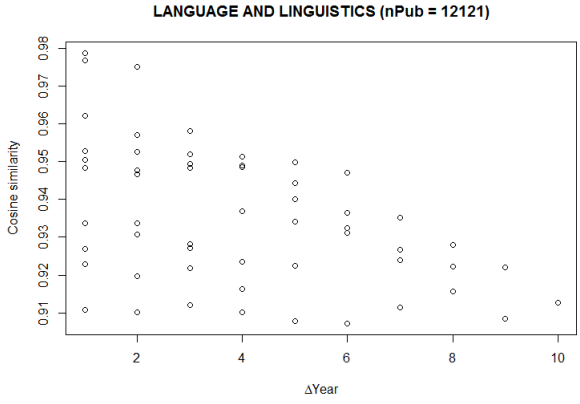

GENERAL AND INDUSTRIAL ENGINEERING (nPub = 12664)
Cosine similarity
ΔYear

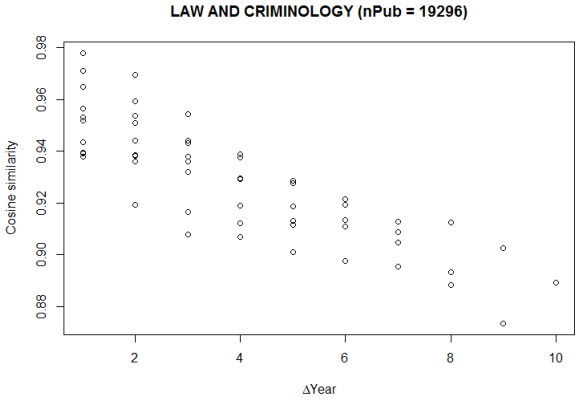

HEALTH SCIENCES (nPub = 101891)
Cosine similarity
ΔYear

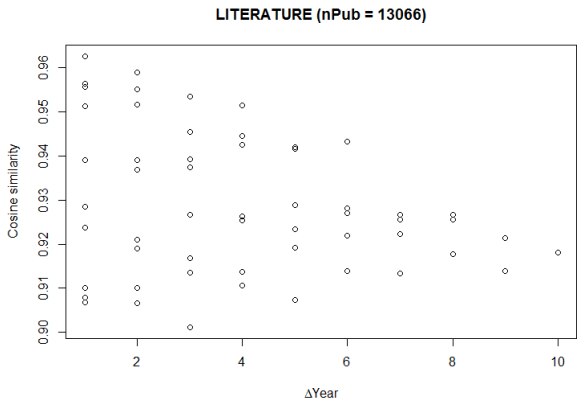

HISTORY, PHILOSOPHY AND RELIGION (nPub = 50461)
Cosine similarity
ΔYear

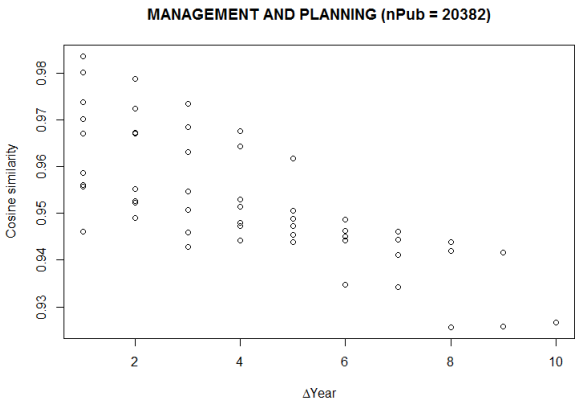

INFORMATION AND COMMUNICATION SCIENCES (nPub = 16819)
Cosine similarity
ΔYear

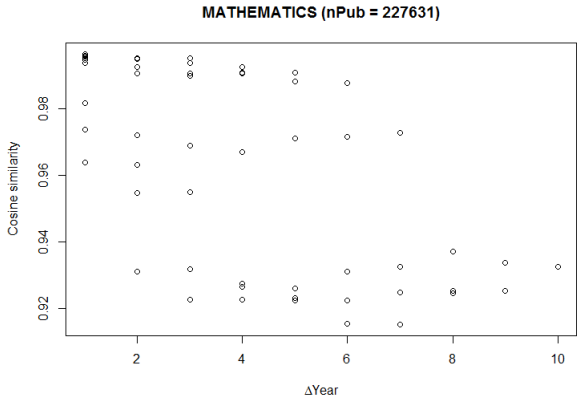

INSTRUMENTS AND INSTRUMENTATION (nPub = 7236)
Cosine similarity
ΔYear

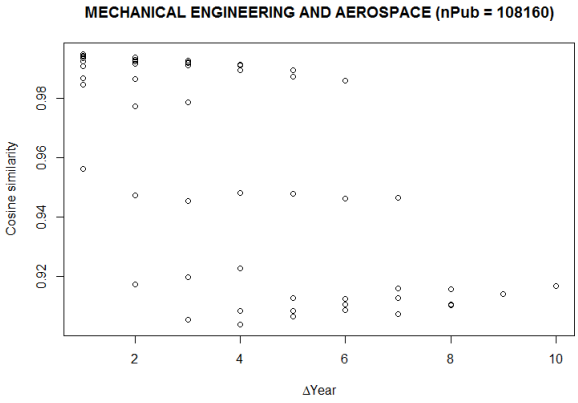

LANGUAGE AND LINGUISTICS (nPub = 12121)
Cosine similarity
ΔYear

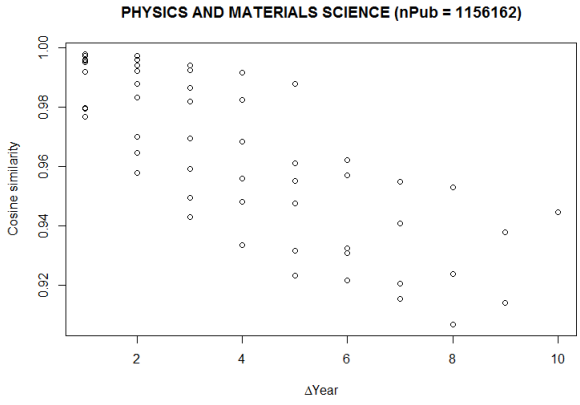

LAW AND CRIMINOLOGY (nPub = 19296)
Cosine similarity
ΔYear

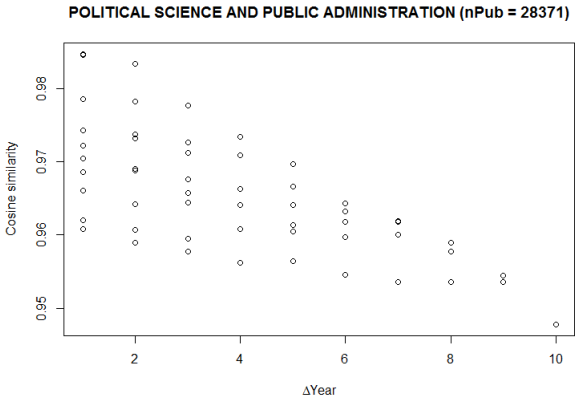

LITERATURE (nPub = 13066)
Cosine similarity
ΔYear

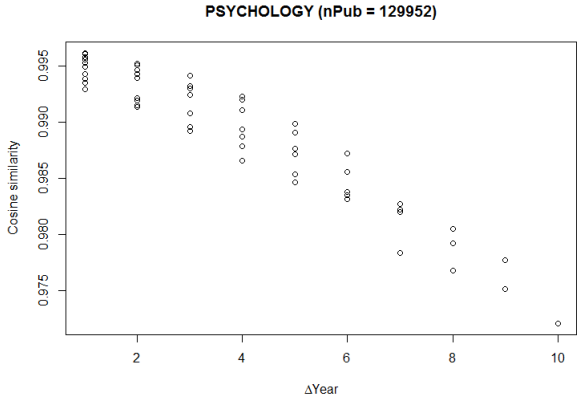

MANAGEMENT AND PLANNING (nPub = 20382)
Cosine similarity
ΔYear

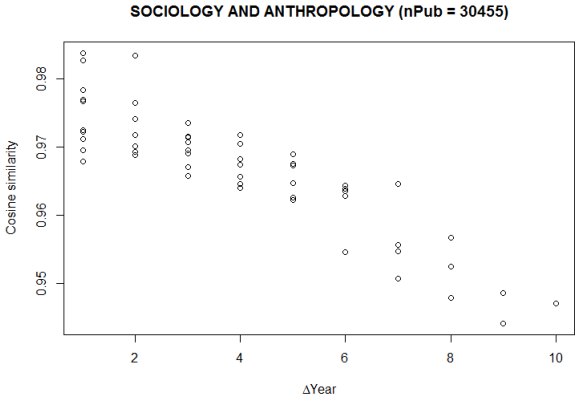

MATHEMATICS (nPub = 227631)
Cosine similarity
ΔYear

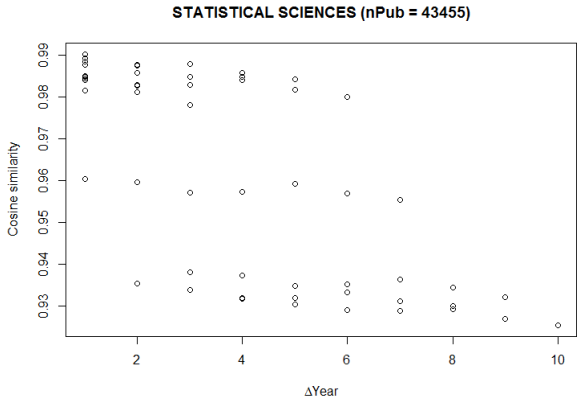

MECHANICAL ENGINEERING AND AEROSPACE (nPub = 108160)
Cosine similarity
ΔYear


PHYSICS AND MATERIALS SCIENCE (nPub = 1156162)
Cosine similarity
ΔYear


POLITICAL SCIENCE AND PUBLIC ADMINISTRATION (nPub = 28371)
Cosine similarity
ΔYear


PSYCHOLOGY (nPub = 129952)
Cosine similarity
ΔYear


SOCIOLOGY AND ANTHROPOLOGY (nPub = 30455)
Cosine similarity
ΔYear


STATISTICAL SCIENCES (nPub = 43455)
Cosine similarity
0.93
0.94
0.95
0.96
0.97
0.98
0.99
2
4
6
8
10
ΔYear

Appendix E: Elaboration on R for data processing

In this appendix we briefly discuss R and its interplay with VOSViewer, the primary software tools we used when processing and visualizing our data and results.

R has several advantages over SQL when it comes to data processing. These advantages prompted us to limit our use of SQL to the extraction of the relevant tables only and made data preparation and analysis almost exclusively reliant on R. While SQL excels as a means for accessing relational data from large databases, the format of the data returned by its queries is restrictive in the sense that it is largely limited to tables and table-like structures, rendering its use limited for processing data in other formats. In contrast, R is a programming language specifically created for statistical computing and gives the user great control over large amounts of data. R offers a wide variety of data structures, primarily vectors, matrices, data frames and lists, in addition to tables, as well as various ways of combining different data formats. R thus allows for a stepwise execution stepwise or algorithmic computations or construct tree-like data structures. A second advantage of R is that it gives the user near complete control over the way data is processed. Given the limited resources available in this research and the relatively large amount of data, the ability to control exactly how and when data processing steps were executed was vital, as was the ability to control which data was kept in memory and the ability to write and read intermediate data to and from disk to free up that memory for further computation.

Finally, R is open-source and third parties can write and distribute their own "packages" through the CRAN (Comprehensive R Archive Network) package repository, containing functions written specifically for types of data or types of analysis. This means that even more complex statistical analysis typically performed using dedicated software such as SPSS is possible in R without too much extra effort when expanding R's default functionality with the right packages. There also exist a large amount of different packages aimed at processing text data in R. Because we can extract our noun phrase occurrence data from the CWTS WOSKB database, their use was not necessary, but had this not been possible, or should we want to refine the noun phrase selection criteria in future research, these packages offer us the tools to expand text data processing. Examples include the RTextTools package (Jurka & Collingwood, 2013) which includes functions for simpler operations such as stemming, and the openNLP package (http://cran.r-project.org/package=openNLP/) which contains R implementations of more complex natural language processing functions of the Apache openNLP toolkit (http://opennlp.apache.org/) developed by the Apache Software Foundation. This latter package is the same one as used in the VOSViewer software package (for detailed discussion see van Eck et al., 2010).

For the visualization of the discipline similarity structures in chapter 4.4, we wrote an implementation of the cosine similarity formula specifically for our data in R, and applied it to each pair of discipline centroids for each year in the sample, as well as for the total sample disciplines. Computing the cosine similarity between each discipline allowed us to generate a network of disciplines and their similarity to each other. We can then use software tools to visualize this network. Our tool of choice is the CWTS-developed VOSViewer (http://www.vosviewer.com/, van Eck & Waltman, 2010, 2011), a frequently used tool in the fields of descriptive bibliometrics and bibliometric mapping (e.g. Leydesdorff et al., 2013; Waltman & van Eck, 2012). This software package is specialized for generating maps based on bibliometric networks based on bibliographic coupling, co-citation or co-authorship, but any sort of network may be visualized as long as it is offered in the proper format; a script was written to format our discipline cosine similarity data in a way that VOSViewer would accept, and the software was used to generate both visual images of the maps as well as VOSViewer map files.