Scientific structures in context : identification and use of structures, context, and new developments in science

Buter, R.K.

Citation

Buter, R. K. (2012, April 26). *Scientific structures in context : identification and use of structures, context, and new developments in science*. Retrieved from https://hdl.handle.net/1887/18707

Version: Not Applicable (or Unknown)

License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden

Downloaded from: https://hdl.handle.net/1887/18707

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University
Repository

The handle http://hdl.handle.net/1887/18707 holds various files of this Leiden University dissertation.

Author: Buter, Reindert Klaas (Renald)
Title: Scientific structures in context : identification and use of structures, context, and new developments in science
Issue Date: 2012-04-26

Scientific structures in context

Scientific structures in context

Identification and use of structures, context, and new developments in science

Proefschrift
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof. mr. dr. P.F. van der Heijden,
volgens besluit van het College van Promoties
te verdedigen op donderdag 26 april 2012
klokke 16:15 uur
door
Reindert Klaas Buter
geboren te Hasselt
in 1973

Promotiecommissie

Promotor:	Prof. dr. A.F.J. Van Raan
Overige leden:	Prof. dr. D.D. Breimer Prof. dr. J. Kok Dr. M. Luwel Prof. dr. R.J.W. Tijssen

Though it should be questioned, whether beauty be not something real, and different from the power of producing pleasure, it can never be disputed, that as surprise is nothing but a pleasure arising from novelty, it is not, properly speaking, a quality in any object, but merely a passion or impression in the sould.

— DAVID HUME, 'A Treatise of Human Nature', SECT. VIII.—*Of Beauty and Deformity*

The typesetting of this thesis was done in ConTEXt.

Voor Mirjam.

Preface*

Let's face it: *science has a problem.* Luckily, it is not some adolescent issue of identity—four hundred years have given science enough resilience to remain self-confident in that respect. Nor is it that the demand for scientific knowledge started to decrease—quite on the contrary, it is needed more than ever if we want keep and improve our standard of living. No, it is a big problem of image which challenges science: a surprisingly large number of people distrust scientific results and cannot connect to the line of reasoning that it employs. Although science is not facing this type of problem alone, but shares it with Christianity, Islam, Capitalism, Socialism, and perhaps even Modern Society as whole, this shared fate provides only shallow comfort. Currently, the image problem is most prominently featured by the debate surrounding research into climate change, which appears divided into camps of 'sceptics' and 'believers'. The importance of this debate and its consequences for science as a whole, is illustrated by a recent episode of the BBC programme 'Horizon' (BBC and Horizon, 2010) in which Nobel laureate in Medicine (Nobel Foundation, 2001) and current president of the Royal Society, Sir Paul M. Nurse, investigated the climate controversy. Still, climate research is not the only topic which is met with criticism: evolutionary biology, genetic engineering, vaccines, and nuclear power all share this fate.

We can think of many reasons why these topics, or even science as a whole, are regarded by some with suspicion. In our opinion though, the most important factors which contribute to the distance felt between those who have a background in science and those who have not, are the *complexity* of scientific knowledge and the nuanced, abstract mode of reasoning that is familiar to many researchers, but alien to many others. Moreover, scientific reasoning is designed for discussion and allows for the assessment of different and even opposed opinions. As a result, it is for outsiders not immediately apparent who to consider *the* authority on a particular subject. Added to that, many researchers are not used to expressing themselves in simple, straightforward phrases, as they are acutely aware of the loss of information and relevant *context* that accompanies such a style. Failing to create such links between how many people *view the world* and the relevant scientific thinking not only adds to the problem of image, but also hinders the application of science. For example, in order to match scientific research and sustainable development, Clarke (2002, p.814) writes that "[a]voiding [...] communication problems will be key to the success of the

---

* Originally, the title of this section was not "Preface", but "Ground control to Major Tom", a line from the famous David Bowie song 'Space Oddity', the first track from his 1969 eponymous album. One of the themes in the song is the lost communication between 'Ground Control' and 'Major Tom' and the apparent inability of both to change this. We used it as a metaphor to refer to the apparent disconnection between science and society, leaving it to the reader whether 'Ground Control' refers to science or society.

strategies being developed.”

However, one could also hold that all the resistance, scepticism, criticism, and the plea for clear communication, is actually a (perverse) acknowledgement of the important role science plays in our society. When complex problems emerge, society usually looks for ‘scientific’ approaches to solve them. Vice versa, challenging societal problems will usually attract attention of researchers and may even become research areas when given enough time, funding, and importance. An example is ‘sustainability science’ which developed from the interest in ‘sustainable development’, and the roots of which we investigate in a publications that is part of this thesis.

So, can the tools developed as part of the research in this thesis help to solve science’s image problem? Or do the analyses shed any more light upon the causes of the communication problem between science and society? We do not think so; at least, not directly. However, we have developed generic tools and techniques to help analyse science and thus may help understand research and their results, as captured in scientific publications. Added to, we have shown how large-scale analyses can reveal new developments, intriguing patterns, and show how the knowledge base of a newly developing field evolves. Moreover, one of the basic themes underlying our research was to increase the applicability and usability of the results of quantitative science analyses. For even in science, every participant has his or her own, individual, unique understanding of the topics under investigation. And every field is characterised by a kind of overlap of all these individual views, which we call the common ground in this thesis. Having the ability to take into account or use those views, or a common ground, is part of the challenge we have taken up in this thesis. Doing so, may allow us to understand science better, because we can provide quantitative results in a context familiar to researchers. And yet, this could be a start for a better understanding between science and society: for asking others to understand us, starts by understanding ourselves.

Dankwoord

Geen enkel werk van betekenis komt tot stand zonder dat anderen daar op één of andere manier aan hebben bijgedragen. En ook ik heb het geluk veel mensen om me heen te hebben die aan dit proefschrift hebben bijgedragen, hetzij direct of indirect.

Bijzondere dank gaat uit naar mijn ouders, Goos en Gerda, die altijd vol vertrouwen geduldig bleven geloven in een goede afloop.

Daarna komen natuurlijk Yaël en Amarinthe! Al die avonden en dagen dat ik druk bezig was met schrijven heb ik jullie (ook) gemist. En ik wil jullie alletwee bijzonder danken voor jullie enthousiasme als ik een weer eens een boekje *voorlas* in plaats van er aan een te werken!

Ook Jeep en Grada, Maroeska, Jordy, Frank en Rieke: fijn dat jullie er waren om te luisteren en te motiveren. En dat geldt voor al mijn familie en vrienden: dank voor jullie steun, geloof in een goede afloop en voor het aanvoelen wanneer het *geen* goed moment was om te praten over de stand van het onderzoek.

Vervolgens mijn (voormalige) collega's van het CWTS. Om te beginnen wil ik mijn promoter prof. dr. Ton van Raan bedanken, voor al het leeswerk en bovenal voor het rode pennetje: daaruit viel, als je eenmaal over de schok van het bloedbad heen was, toch bijzonder veel te leren. Daarnaast moet ik ook Martijn bedanken, kamergenoot gedurende zo'n lange tijd: onze on-topic en off-topic discussies waren altijd leerzaam en het was een heerlijk gevoel samen op te trekken in onze pogingen om nieuwe dingen tot stand te brengen. Uiteraard moet ik Ed ook bedanken, co-auteur van mijn eerste publicaties en mentor gedurende lange tijd bij het instituut: dank voor het wijzen in de goede richting. Maar ook Thed, Bert, Robert, Cornelis, Marc, Suze, Clara, Peter, Erik, Ton, Rodrigo, Paul, Nees-Jan, Henk, Maria, Ludo en Christine: allemaal bedankt voor de interessante, leuke en uitdagende dingen die we samen hebben gedaan.

Daarnaast de mensen in binnen- en buitenland waarmee ik de afgelopen jaren heb mogen samenwerken in nationale en internationale projecten: het gaat te ver om jullie allemaal bij naam te noemen, maar bij deze wil ik jullie allemaal zeer hartelijk danken voor alle goede ideeën en de vruchtbare discussies..

Als laatste wil ik de persoon bedanken die de *allergrootste* bijdrage heeft geleverd aan dit proefschrift, zonder daarbij ook maar een letter ervan te schrijven. Zonder haar steun, koppig optimisme, *mental coaching* en bovenal pragmatische houding, had ik het nooit kunnen volhouden en afmaken. Lieve Mirjam: *de top is bereikt!*.

Contents

Part I

General Research Background and Bridge to Part II

1

Introduction and general background

1.1 On science

Science is a complex phenomenon with many facets: cognitive, educational, social, political, economical, geographical, philosophical, and many more. It touches upon many aspects of modern life and involves many different types of people; either directly or indirectly. It is also a highly productive system: each year, apart from the thousands of students, hundreds of instruments, prototypes, and experimental medications, it produces many publications; world-wide, there are currently over one million published each year in peer-reviewed science journals. And although budgets are tightened, governments and organisations still invest heavily in the development of scientific knowledge. For example, even though the optimistic objectives of the Barcelona agreements appear difficult to meet[1], the EU still invests over 50 *billion* euro[2] in the Seventh Framework Programme (FP7) which runs between 2007 and 2013. In the last part of the twentieth century, scientific knowledge has established itself as a standard, enabling a global market for both specialised and applied knowledge. So, even though the nature of science already an interesting subject to study, there are many additional reasons why it is both interesting and rewarding to understand how research is organised, who is excellent at what, and what current developments are.

However, the complexity of the subject forces us to choose what facets to study. The research we present in this thesis focuses on the cognitive facet, which means that we are interested in research topics and the description of such topics. Still, the social facet is (at least implicitly) present as well, simply because people are central to science. Additionally, our analyses are mostly quantitative as opposed to purely qualitative, which means that we base our conclusions on numbers, as opposed to interpretations of lab texts or interviews, for example. The numbers we use for our analyses, are drawn from scientific publications from peer-reviewed journals. The use of such publications has at least two important advantages: first, for a large part of modern science this type is the *de facto* unit of publication, which means that we

---

[1] As we can infer from for example the statistical annex at `http://ec.europa.eu/research/era/pdf/statistical-annex-stc-report_en.doc`.

[2] This number can be found on the CORDIS website, `http://cordis.europa.eu/fp7/budget_en.html`

can cover a broad area with our analyses; second, the peer-review system establishes a basic level of quality on which we can built. Of course, this also means that we do not extensively cover areas of science where other publication types are more common, like computer science where much is published in conference proceedings, or the humanities and social sciences where books are very important. Also, we do not cover the knowledge embedded in (peer-reviewed) bodies of grey literature, such as countless numbers of reports, master theses, and PhD theses. Still, much of what we used, developed, and applied in the research for this theses, is generic and much could be readily adapted for other publication types.

We noted that research topics are a central interest in our analyses. We regard research topics as collections of knowledge, including concepts and research questions. Such things we cannot actually hold in our hands or store in a database: they are *intangible*. So, we necessarily investigate matters by using things we *can* store, the *tangible* representations of concepts or questions in scientific publications, expressed by things such as words, affiliations, author names, or references. By combining and aggregating these tangible items, we try to figure out how they are structured, how these structures develop, and then infer what this means in the larger context of the research we investigate. As we already noted above, such analyses have both academic and practical value. So, although the tools and analyses we describe in this thesis have been published in established, scientific journals, our work has not been a detached, academic exercise, as most of what we have developed had its origins in actual (policy) demand.

1.2 Summary of Part I of this thesis

In this first part of our thesis we introduce the general background of our work. We have divided this first part two chapters: the "**Introduction and general background**" you are reading now and the "**Bridge to Part II**". This introduction is divided into four sections:

1. "**Knowledge, meaning, and cognition**" (**Section 1.3**);
2. "**Science, output and analysis**" (**Section 1.4**);
3. "**Change and developments in science**" (**Section 1.5**);
4. "**Visualisation and maps**" (**Section 1.6**);

Together, these four sections comprise aspects central to the work we present in Part II of this thesis. We start in **Section 1.3** with the question how knowledge can be communicated and understood. We investigate how we should understand the knowledge in the minds of people and captured in the language and references used in scientific publications. This is an important topic, since a number of problems we address in our papers arise from this fundamental gap between tangible and intangible aspects. After that, we move on to the topic of science, scientific output, and the (quantitative) analysis of science in **Section 1.4**. We essentially regard science as a system in which people do research, write about the results of their work in publications, while building on and expanding the work of others in the process, and who are organised in institutions that provide the necessarily facilities for them to do this. The scientific discourse that is central to this system is organised in specialised areas of research, just like the people and institutions in science. In **Section 1.5** we explore the background on new developments in science, again from the perspective of scientific publications. We explain that new research developments have different phases: birth, spread, maturing, and institutionalisation. We also explain that emerging (new) research faces some inherent headwind due to the way science works, including the importance of prestige, and the necessary cautiousness in the peer-review system. Many academic contributions in our field are analyses of newly developed areas. Our contribution, however, has been in finding such emerging areas, besides investigating the origins of such areas. More specifically, we have contributed to the search for converging research, which are developments where two previously separated fields start to connect. Such interdisciplinary developments require even more effort, as researchers are required to become familiar with a new environment and its own jargon, important publications, as well as existing social structures. We also contributed to the investigation of a particular emerging field (‘sustainability science’), which is the result of a societal and political push to find sustainable ways for development. Finally, in **Section 1.6** we turn to visualisations, such as networks and maps. Visualisations can help locate influential publications or actors, relations, communities, and trends. As such, they are important tools to comprehend complex bodies of knowledge, such as captured in scientific publications. The structure displayed by the map can be used for domain analysis, scientometric analysis, or trend analysis. One of our contributions is the design of a tool which can perform multiple tasks, such as the retrieval of detailed bibliographic information for an element on a map, as well as the display of numerical data in charts, as labels, or as colours. Additionally, we used maps and structures to filter terms and to bridge intangible concepts and tangible terms in science. After we have introduced the general background of our research, we summarise in **Chapter 2** our contributions in Part II (Chapters **3** to **9**) and show how these contributions fit into the general background we describe in **Chapter 1**.

1.3 Knowledge, meaning, and cognition

Science is *literally* knowledge: from the *Oxford English Dictionary* we learn that the word is rooted in the Latin verb *scīre* (‘to know’) via the present particle *scientem* (‘knowing’)[3]. There are a number of notable characteristics of ‘scientific’ knowledge.

---

[3] The present particle in English has a form identical to the gerund. To understand the difference, we can think of the gerund mostly as a noun (‘His power was *knowing* men.’), whereas the present

- One important characteristic is the *type* of knowledge that is considered 'scientific'. For example, the type of knowledge collected in astronomy and chemistry is considered scientific, while that in astrology and alchemy is considered occult and not scientific[4].
- Scientific knowledge is more than (passive) facts, but includes *methodology and tools*, i.e. rules and techniques on how to combine, aggregate, select and *create* knowledge. Examples are inductive reasoning, clustering algorithms, and models created for complex systems like local weather and the global climate.
- The body of knowledge *evolves and adapts*, and parts may become stale or obsolete. The hypothesis that light exists of tiny particles which travel in a substance called 'æther', for example, was at one time considered seriously by many respected physicists (Chalmers, 1999).
- Scientific knowledge has a strong sense of *locality*, i.e. where it was *created*, although *application* of knowledge is not restricted to where it originated. For example, as noted in the introduction of Rinia et al. (2002) the concept of the electron spin[5] was developed in (quantum) physics, but a well-known *application* is in the MRI[6] scanner, an important instrument in the neurosciences.
- Even the body of scientific knowledge that belongs to a single area is already so complex and large, that to *access* or understand it, a person requires considerable training.
- Finally, not all contributions to the scientific body of knowledge are considered equal: the research community considers some to have more *impact* (be more fundamental, more insightful) and thus be more important, than others.

Although the individual characteristics (type, methods and tools, development and obsolescence, locality, access, and impact) are not enough to differentiate scientific from what we could call 'general' knowledge, when taken together, they clearly indicate a special flavour. Also, in the analysis of science, these characteristics play an important role, which we will also try to make more apparent in the coming sections of this first part of our thesis. First however, having established that knowledge is central to science, we will continue with the question how knowledge, which has no substance, can be communicated and understood, as well as captured in both publications and the human mind.

---

particle modifies something else ('He was surrounded by powerful, *knowing* men.')

[4] Although it is well known that Isaac Newton, a founding father of modern science, showed profound interest in alchemy (Figala, 2002).

[5] The concept of the "electron spin" was first published by Uhlenbeck and Goudsmit (1925), and can be considered an important Leiden contribution to quantum physics, as both Goudsmit and Uhlenbeck were students of Paul Ehrenfest during his Leiden period, which lasted from 1912 until his unfortunate death in 1933.

[6] MRI stands for Magnetic Resonance Imaging.

1.3.1 Communication and meaning

Communication is the mutual exchange of information in the form of encoded messages. In this thesis we are interested in communication where a message is (part of) a scientific publication, and senders and receivers are either researchers or other individuals with an interest in science (such as a policy makers). Evidently, in science there is usually no point in exchanging singular messages, because participating in a larger discussion is more important than merely presenting self-contained results. Having a discussion implies referring back to previous messages, but also referring to the context in which arguments should to be understood. As a result, scientific publications are highly structured and contain many links to other publications as well as other bodies of knowledge. Hence, we have one of the reasons why there is so much structure in science: the scientific discourse demands it.

The exchange of messages (writing and reading publications) only makes sense if receivers are able to assign *meaning* to the code[7] of the message. The most general (or abstract) study of meaning is *semiotics*: "the study of sign systems". In Saeed (2003) two similar definitions are given, attributed to two central figures in semiotics: one by C.S. Pierce, who calls it the study of "signs, objects, and interpretants", and one by De Saussure, who calls it the study of "signs, signified, and signifiers". We will use the terminology by Pierce in this thesis. The three elements can be placed a triangle, the so-called *sign triad.* In **Figure 1** we illustrate how we use this triad in this thesis to explain the relations between sign, object, and interpretant. In the figure, we have labelled the relations from the perspective of the (human) interpretant who *sees* a sign, and *resolves* it to a *known* object. The relation between sign and object can range from concrete to abstract. For example, in a painting an apple can concretely refer to the actual fruit, but also abstractly to the fall from grace of mankind. Signs are encoded in some coding system, and (this encoding) is the only 'tangible' element in this triangle.

Closely related to the concept of sign, is the concept of *proxy.* We consider a proxy to be an object for which a sign is readily available, and that can 'stand in' for some other object, for which a sign is *not* available. A good example of a proxy in science analysis is the use of a set of journals as a proxy for 'field': a direct representation of a science field would be difficult, due to the fuzzy character of the concept.

Despite its general character, semiotics has been applied in science analysis. One example is Cronin (2000), who used the sign triad to illustrate how references, acknowledgements, citations, and impact (citation counts) differ. His use and interpretation of the sign triad has been an important influence on the way we use it here.

Another example is Mai (2001), who used semiotics to explain that indexing terms are not signs which refer objectively to particular subjects, but are the result of a process of interpretation.

---

[7] Understanding a message is, however, completely different from *accepting* it. For example, economists have long before the current financial crisis, warned about the potentially damaging effects of the U.S. housing bubble, or the Greek deficit spending. Although such messages have been written and probably also been read *and* understood by influential people, this was apparently to no avail.

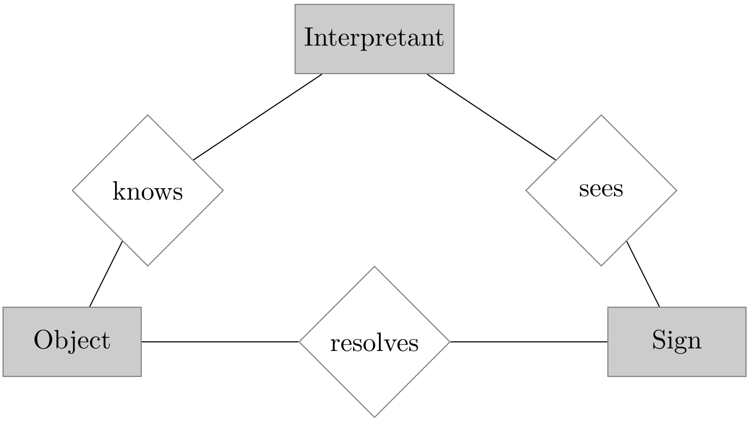


Figure 1 Illustration of our use of Pierce's sign triad, drawn from the perspective of the (human) interpretant.

1.3.2 Meaning and language

Depending on what is studied, semiotics has several specialisations (Saeed, 2003, p.17), which are increasingly abstract and intangible:

- *syntaxis*: the study of the relation of signs to each other;
- *semantics*: the study of the relation between signs and the objects to which the signs are applicable;
- *pragmatics*: the study of the relation between signs and their interpreters.

A scientific publication consists of signs in different encodings (text, figures, and tables) referring to (abstract) objects. As (encodings of) signs are the only tangible elements, they are also the only things which can be stored: either on paper or in a database. As a result, quantitative analyses of science necessarily start at the syntactical level (dealing with text, authors, publication year, and other items of publications), after which we can lift the results from the syntactical to the semantic level, by interpreting the results. The interpretation links objects to the results and how these links fit into a larger whole or *context*. We will get back to context in **Section 1.3.4**.

Natural language is the most important way in which a researcher tries to convey a message, and text is the most used encoding. Language is a highly flexible and adaptive vehicle, but is also somewhat of a mixed blessing. An important problem is that of homonyms and synonyms: for one object, we can have multiple signs and *vice versa*. For example 'cloud' may refer to either the physical phenomenon in the sky, or to a type of distributed computer platform; and 'impact' is commonly used in bibliometrics to refer to citation performance, but may refer to other objects as well, for example from astrophysics. To be clear on what authors mean, they have to add more information. A sentence contains words with different functions, like verbs, nouns, adjectives, and adverbs. Much of the meaning a sentence tries to convey is captured by nouns. A noun can also appear in a group of words affect the meaning of this noun: such a group is called a Noun Phrase (NP, plural NPs, see Voutilainen (1995), as well as the introduction of Noyons (1999)). An NP has a more specific meaning than the noun and is thus an important tool for authors to give more information: instead of 'impact' they can write 'citation impact' or 'asteroid impact', depending on their intention. Another problem with language is that there are many flavours and that only a few are chosen as a vehicle for scientific discourse; for authors and readers who were not raised in such a language, this has obvious consequences for both conveying and understanding messages.

Added to that (and barring trivial publications) the set of signs in an average publication is large. So, if individual signs in publications can already cause interpretation problems, then the *combination* of many signs makes it even more difficult to be certain that the *intended* message of an author will be understood without distortion by most interpretants (readers). Not only that, but the perceived informational value of a publication is *dependent* on the reader and for what purpose the document is read (Buckland, 1991). In other words: a reader adds information while reading a publication. The reason for this is that we have default assumptions (Hofstadter, 1985, Chapter 7), which can be thought of as preferred shortcuts to objects. These assumptions guide our understanding and can, when enough context is given, also fill in blanks, by creating extempore objects. This is often done subconsciously, so readers will not even know the difference between understanding and filling in. An author of a scientific publication has little means (beyond the text, figures, and tables) to guide the interpretation process in a reader, although he can assume that a reader may possess certain knowledge.

1.3.3 Cognition

If we interpret a publication as a set signs encoded as text, figures, or tables, which can be stored on paper or electronically. However, this is not possible for the objects to which these signs refer. At most, these objects can be considered as being are encoded and stored in the minds of people and it is therefore interesting to briefly look into cognition.

The representation used to store information in a human brain is difficult to assess, but it has long been established that "barring decisive physiological data, it will not be possible to establish whether an internal representation is pictorial (visual) or propositional (verbal, conceptual)" (Anderson, 1978). Still, the way people acquire knowledge may make a difference, and we can differentiate for example between procedural knowledge and survey knowledge (Thorndyke and Hayes-Roth, 1982). Although all information enters the brain via the senses, there is no clear-cut connection between the stimulus and the information. In Landesman (1997) a distinction is made between 'direct realism' and 'representative realism'. Direct realism is the direct sensory input, like touch, taste, smell, hearing, and seeing. However, it is not always possible to directly connect sensory input with information. Landesman uses the example of a hot potato, which, upon touching, carries a sensation which we interpret as 'hot', but the interpretation 'hot' is not the same as the pain we experience directly. Instead, background knowledge on the what kind of sensory inputs *represent* or correspond to 'cold', 'warm', and 'hot', was combined with the knowledge that a potato was touched, in order to deduce that this particular specimen had to be a 'hot potato'. Extend this to science and we could say that "scientific enquiries are those which go beyond what is available to simple reflection and our original understanding" (*ibid* p.15). Of course, this observation has a direct relation to the sign triad of the previous section.

Although useful, the distinction between 'direct' and 'representational' realism is rather crude. In De Mey (1982) a more fine-grained classification is introduced which is used to classify philosophies of science. As the scheme was originally used to classify different stages in information processing[8], it is straightforward to recast it to fit different levels of scientific knowledge and understanding. The classification consists of four levels, each subsequent level adding additional complexity to the former.

- **The monadic level** is the level of basic concepts or facts, without relations between these facts. Examples are concepts like 'spin', 'tree', or 'electron'.
- **The structural level** is the level of simple syntactical rules which can be used to combine and transform facts. An example would be 'electron spin'. However, if a rule fits multiple facts (or none, for that matter) or has multiple outcomes, then at this level there no way to decide which of the applications or outcomes are correct. For example, we cannot judge whether 'tree spin' makes as much sense as 'electron spin'.
- **The contextual level** adds meaning and knowledge on what rules can be applied onto what facts. For instance how the electron spin fits in the larger context of quantum mechanics.
- **The cognitive or epistemic level** introduces a *world view* of the researcher, which is a selection and aggregation of the things experienced and learnt over years. For example, how (and if) quantum mechanics fits in how we observe the world at the non-quantum scale.

There are similarities between the last three levels of this classification and the three semiotic theories we mentioned above (syntaxis, semantics, and pragmatics). The difference is that the above classification is expressed in terms of internally coded facts or concepts, and not in terms of externally coded signs.

---

[8] Which was an adaptation from a classification Donald Michie devised in the context of Artificial Intelligence.

1.3.4 Context and common ground

It is said that “context is everything” and we have already used the word a number of times in this chapter. Providing context is an important instrument to align the (application of the) individual world views of readers, in order to guide the interpretation of a message (in a publication) and have a worthwhile scientific discussion. In a sense, such discussions are held on the basis of what Stalnaker (2002) calls *common ground*: the *presumed* background knowledge of a publication or (by extension) of a research area or field. There are a number of ways to express or codify common ground in scientific publications. One is to use certain words that refer to a central concept or technology that does not need additional qualifications in a publication. We refer to such words as ‘jargon’**[9]**. Highly cited publications can be used as well (Small, 1978). Even authors and institutions can refer to a body of knowledge or context, a practise that is more common in the social sciences and the humanities than in the natural sciences. For example, we can speak of the ‘Chicago School’ or the ‘Wiener Kreis’, without having to explain their stand on certain sociological or philosophical issues. As such, these references are closely related to Kuhn’s ‘exemplars’ (see **Section 1.5.2**), and we will use this as the general term for references to knowledge in this thesis. Also, at some point certain parts of a common ground can become so accepted, that authors are no longer required to reference them. This includes much of the generic knowledge that is taught while following an education at a university, such as integration and differentiation in calculus, which was once considered expert knowledge. Also basic publications, such as Einstein’s papers on special relativity, are hardly cited anymore because they are now part of an accepted common ground. Moreover, to know and understand accepted parts of a common ground is why it may take years of training to use (let alone contribute to) a particular body of scientific knowledge.

1.3.5 Summary

To summarise, when analysing science, we always have to deal with two levels: the tangible representation and the intangible intention. By introducing semiotics and its concepts of sign, object and interpretant, we have tried to explain how knowledge, communication, meaning and cognition relate. Signs are the vehicles of communication, meaning is to be understood in context of the represented objects and the person who interprets their meaning. Cognition is the ability that allows people to attach meaning, and built up a set of internal structures at different levels of complexity, the

---

[9] Interestingly, the Oxford English Dictionary, notes as the first meaning of ‘jargon’ the “[...] inarticulate utterance of birds, or a vocal sound resembling it; twittering, chattering [...]”. Only the sixth description resembles our use: “[...] unfamiliar terms, or peculiar to a particular set of persons, as the language of scholars or philosophers, the terminology of a science or art, or the cant of a class, sect, trade, or profession”. Essentially, this would make our use of jargon itself an example of jargon. Fortunately, the Merriam-Webster Dictionary gives “the technical terminology or characteristic idiom of a special activity or group” as second description, in line with our use of the term.

end result of which we call a world view. To have a worthwhile scientific discussion, participants need to be able to refer to a common ground, which is the shared (and implicit) background knowledge in a field. Common ground can be encoded in terms of specific words or phrases, references, or even names of people or institutions. Indicating the context in publications is one of the mechanisms that structures scientific discourse.

1.4 Science, output and analysis

Let's step back and look at the system that enables 'doing science', as performed by researchers. Tijssen (1992) distinguishes three main aspects of the system:

- the transfer of knowledge ('process'); and
- the 'size' of scientific activities, measured by the output in publications ('product');
- the social and cognitive networks of science ('structure').

Note that these aspects are interdependent: the transfer of knowledge is mostly done through research publications that are read, produced and published in and by the social and cognitive networks. Also note that some aspects are tangible and some are not. Buckland (1991) distinguishes between:

- *information-as-knowledge*, the intangible 'being informed' that is result of the consumption of knowledge; and
- *information-as-thing*, the tangible 'data or document' that is the result of the production of knowledge.

In the quantitative analysis of science, we always base our analyses on tangible elements, and from that, we infer things about the intangible ways in which knowledge and research shape science. In the previous section we focused on the the intangible aspects related to the process of communicating knowledge. In a next section we will focus mostly on (visualising) structure. In this section though, we will focus on the scientific system and one of its most important products: the scientific publications.

1.4.1 Collections of scientific output

We can only use the tangible products of the scientific process if they are collected somewhere. There are a number of such collections we can draw data from. For example, many departmental and project websites contain pages that list publications published by that department or as a result of that project. Sometimes the lists also contain links to the full-text versions of the publications. The information on these websites can be collected in repositories, such as CiteSeer[10] and Scholar from Google[11] which allow users to search for publications. Additionally, there are repositories to which researchers submit their articles, such as arXiv[12] (pronounced as 'archive'), or institutional repositories, which contain both the actual publications and the meta-data of these publications, such as authors, title, means of publication (article, book), or publisher. Bibliographies on the other hand, only contain the meta-data[13], sometimes supplemented with an abstract. Some bibliographies are web-based, while others are restricted to a local intranet. The web-based bibliographies are either publicly available or for a (often quite expensive) fee. The freely accessible bibliography DPBL[14], which collects publications related to computer science is an example of a freely available web-based bibliography, just like PubMed[15], which collects publications in the life sciences. On the other hand, Inspec[16], which collects publications in the field of science and technology, is usually only available on restricted intranets or to a restricted set of domains. Although bibliographies do not contain publications, they can provide a link to the location where the publication *is* available. Finally, there are the general bibliographic databases. The two most important ones are Scopus[17] (published by Elsevier) and the Web of Science[18] (published by Thomson Reuters as part of the Web of Knowledge). These general bibliographic databases cover a broad area of science, much broader than the more specialised bibliographies. Added to that, they also store the references of publications. The meaning of references can be viewed in different ways (Moed, 2005a, Chapter 15). One view of is that citations are "judgements, or 'votes' of colleague-scientists" (Van Raan, 1996, p.399), while another view (already noted in **section 1.3.4**) is that citations are 'exemplars' or 'pointers' to knowledge, as promoted by Small (1978). In line with this last view, the set of cited publications of a field, is also called the *knowledge base* of that field. Both views are equally important in the analysis of science to serve as a basis for citation-based assessments of performance and quality.

The mirror image of a reference is a 'citation' (Wouters, 1998, p.233): a citation is the act of *being* referred to. So, if publication P refers to publication Q, then Q receives a citation. Alternatively, we can say that P *cites* Q or that Q *is cited* by P. Also, when two publications Q and R are cited together by P, then Q and R are *co-cited*, while if P and M are both citing Q, then P and M are *bibliographically coupled*.

Most data we use in publications in this thesis are extracted from the Web of Science (WOS). Being a general bibliographic database, it indexes[19] a significant

---

[10] `citeseer.ist.psu.edu`

[11] `scholar.google.com`

[12] `arxiv.org`

[13] Although web front-ends usually contain a link to the actual scientific publication.

[14] `dblp.uni-trier.de`

[15] `www.ncbi.nlm.nih.gov/pubmed`

[16] `www.theiet.org/publishing/inspec`

[17] `www.scopus.com`

[18] `www.webofknowledge.com`

part of the scientific output published in peer-reviewed scientific journals from 1980 onwards. CWTS[20] has licensed this database and used it to create a system that is fully dedicated to perform general bibliometric and scientometric analyses. The database contains meta-data of publications, such as title, affiliations, keywords, and journal, as well as references and abstract. Many references are to publications that are also covered by the WOS, but because references as given in publications may contain small errors or ambiguities, CWTS has invested much effort in reliably finding the referenced publications. The database has also been enriched with additional meta-data regarding institution types, and many author and affiliation names have been disambiguated and normalised (Moed et al., 1995).

1.4.2 Metrics for scientific output

The research area that deals with calculations and data manipulations in the most general sense is called *informetrics*. The description of the scope of the most important journal of this field, the *Journal of Informetrics*, notes that it is kin to other metrics fields, such as econometrics and biometrics. In this field there are two notable specialisations, *bibliometrics* and *scientometrics*. The first is the study of science based on publication data (usually, although not necessarily extracted from a bibliographic database), while the second uses metrics derived from any type of data to analyse science. The difference between the two is diffuse and although bibliometrics is a specialisation of scientometrics, publication data are very valuable and are often the only reliable data available, making both in such circumstances more or less the same type of study. Since the publications in this thesis are mostly based on data from the WOS database, they can thus be regarded as bibliometric papers.

A central activity in bibliometric analyses is the calculation of indicators. Indicators are metrics which measure performance of actors in some aspect, possibly restricted to a specific factor such as topic or field. For example, a citation indicator expresses performance in 'being cited', and an output indicator expresses performance in 'production of scientific publications'. Much work has been done over the years to create good indicators, as they are essential tools in the analysis and assessment of science. Over the years CWTS has used several indicators, such as the number of publications (P), the number of citations without self-citations[21] (Cx), or the average number of citations (CPP). Very importantly, citation counts cannot be compared

---

[19] In this thesis, when referring to items in a database, we arbitrarily use "contain", "collect", "cover", "include", or "index".

[20] The 'Centrum voor Wetenschaps- en Technologiestudies' (CWTS, which is Dutch for 'Centre for Science and Technology Studies') is a leading institute on the quantitative study of science and technology, part of the Leiden University in the Netherlands.

[21] Authors use self-citations to promote their own research, which is a different activity than for example giving credit to a publication; added to that, self-citations could also be abused to artificially inflate citation counts. For both reasons, self-citations are usually removed from the calculation of citation indicators.

over fields, because the average number of references can vary significantly across fields. For example, the average mathematics paper from 2009 in the WOS has 9.3 references, while the average biochemistry paper has 42.9 references. So, before they can be compared, citation counts must be normalised relative to the field in which the publications for which they are calculated were published. A field-normalised citation count that we have often used is the CPP/FCSm (also called the 'crown indicator'), which is the number of citations divided by the citation mean of a field, i.e. for a set of P publications, the CPP/FCSm is $\sum_i^P c_i / \sum_i^P f_i$, where f_i is the citation mean[22] of the fields in which publication i is published. Another well-known indicator is the h-index[23] (Hirsch, 2007), which expresses the number of publications n out of N that have at least n citations, while the other $N - n$ publications have no more than n citations.

An indicator gives a single number, a 'point estimate' (Chaudhary and Stearns, 1996) for a set of publications. We can add more detail by splitting publications into smaller sets, for instance according to their publication year. However, although this shows the development of an indicator over time, we still have no understanding how representative this estimate is for publications in a given year. Knowing such variation could for instance be important when ranking research groups by sorting on (point estimates of) an indicator: if estimates are close and overlap when variation is taken into account, then this could indicate that the rank of these groups is (approximately) the same. A way to estimate variation within a set of indicators, without assumptions about the underlying distribution of the indicator, is to use non-parametric bootstrapping (Efron and Tibshirani, 1993). This method can be used to calculate an estimate of the sampling distribution of a statistic, by repeatedly re-sampling a set and calculating point estimates. The variation of the point estimates can be considered an estimation of the variation of the statistic. Bootstrapping has successfully been used in for example the analysis of costs in medical trials (Barber and Thompson, 2000), phylogenetics (Hillis and Bull, 1993), and economics (Simar, 1992), but its use in bibliometric analyses has thus far been limited. In spite of that, we applied it successfully in one of the publications that is part of this thesis, to differentiate between point estimates of impact in groups of publications that did or did not include non-alphanumeric publications.

1.4.3 Analysis of scientific output

We can analyse either in a descriptive fashion (top-down) or in a evaluative fashion (bottom-up) (Van Leeuwen, 2004). If we analyse top-down, we start with a set of publications and extract and assess the structures within the set. And if we analyse

---

[22] Recently, there has been some debate on the validity of calculation of the crown indicator, which has led a revision in the way CWTS carries out field normalisation (Van Raan et al., 2010).

[23] Although attractive due to the simplicity of this indicator, Waltman et al. (2011) note that from a mathematical point of view, this indicator is not consistent in ranking researchers and groups relative to each other, and should therefore by used with extreme care.

bottom-up, we use such structural elements to get at a more general description of the set of publications from which the elements are part. An analysis of science will usually require both directions, possibly in an iterative fashion. Also, analyses can have several goals. For example, if goal is to find and promote research which performs above average, then we could compare impact and output indicators of actors with those of other, related actors. As such, Van Raan (1996) notes that it is an important tool in assessment procedures, next to peer-review, the qualitative assessment given by peers of an actor. Another goal could be to find promising research topics, or newly developing interdisciplinary research; both of which rely less on the application of indicators and more on the analysis of structural elements.

1.4.4 Delineation of scientific output

Because our analyses are based on publication data, we have to start by collecting the relevant publications (from the WOS). We refer to this process as the 'delineation' of research, in line with Zitt and Bassecoulard (2006), although others have referred to it as 'delimitation' (De Bruin and Moed, 1993), or 'definition' (Lewison, 1999). The delineation of research is an important and challenging part of a bibliometric analysis. The fundamental challenge is to know the proper scope of an analysis and express this as a set of publications. An important source of information are experts, who may also judge the usefulness of the final result. Of course, this introduces a subjective element in the process, akin to for example the validation of a cognitive map, which we deal with in **section 1.6.6**. Since scope is a fuzzy concept, a delineation is often done in iterations, where every next iteration adjusts the scope and delineation based on information extracted from the results of the current iteration. Another challenge is to choose the basis on which publications are collected, such as:

- journals;
- affiliations or addresses;
- natural language;
- keywords from a controlled vocabulary, or a set of codes from a classification scheme;
- references to important publications.

We will briefly glance over the advantages and disadvantages of these alternatives. Of course, we do not have to use one alternative exclusively, but can also experiment with a combination of alternatives, as for example in Zitt and Bassecoulard (2006).

Journals

Many specialised journals have a specific scope, so this is an attractive option to use as a basis for a delineation. Even more, methods have been proposed to select the 'core' (most central) journals of a specific field (Hirst, 1978 and Egghe and Rousseau, 2002). However, using journals has some drawbacks. The most important is that some prestigious journals like Nature, Science, and PNAS[24] are difficult to include, because they have a multidisciplinary orientation and publish publications from different fields. If publications from these journals would be included, then the delineation would contain both relevant and completely irrelevant publications, which would also influence performance indicators. Excluding these journals, however, could exclude highly regarded contributions. Another drawback is that journals may not be appropriate for interdisciplinary research areas where publications are scattered over a (large) number of journals, none of which can be considered 'core' to the area.

Affiliations

Instead of journals, we can also use affiliations to delineate. By using affiliations, we can select publications independent from journals, so Nature and Science articles can be included. Moreover, if the affiliations are chosen well, then this method is also more specific in selecting only relevant publications, so it can be used to delineate interdisciplinary research as well. However, a drawback is that variations in affiliation names in a database such as the WOS can cause the unintended exclusion of publications. Added to that, the scope of an analysis is limited to a specific set of actors, which is not desirable for more general analyses. De Bruin and Moed (1993) have had some success in science field delineations by creating a classification scheme using cognitive (content-bearing) terms in institutional addresses, such '(department of)' 'health' or 'pathology'. In a similar vein, we investigated in one of our contributions whether a similar approach could be used to find newly emerging research (see also **section 1.5.3**).

Natural language

Words and phrases from titles and abstracts can be used to select publications based on their content. This way of selecting publications has the advantage that it can be made to include or exclude publications independent from where they were published, or by whom. As such, they can overcome the drawbacks of the above alternatives. However, as noted in **section 1.3.2**, natural language is somewhat of a mixed blessing as a vehicle of meaning, as it introduces ambiguity and lack of coverage due to homonyms and synonyms. Moreover, people have a preference for certain words, and 'spontaneously' supplied words to describe a topic can result in up to 80-90 percent failure rates in many common situations, something which is called the 'vocabulary problem' in Furnas et al. (1987). An example from Furnas et al. (1987) is that the action 'to remove a word' can be labelled as a 'change', 'removal', or 'deletion'. Another example is the alternatives for 'delineation' found in the academic literature we mentioned above, which all refer to the same process. As a result, when multiple experts are asked to describe a field, then we end up with a number of slightly different

---

[24] The full title is: Proceedings of the National Academy of Sciences of the United States of America

descriptions; these alternatives must be merged and the result will necessarily describe a broader scope then any individual description. Whether that is an advantage or a drawback probably depends on the type of analysis.

Controlled vocabulary or classification

An alternative to natural language is a controlled language (thesaurus) or classification scheme, both of which are usually maintained by a database publisher. Well-known examples are the MeSH (Medical Subject Headings) in PubMed, the Inspec Thesaurus, and the Journal Subject Categories (JSC's) in the WOS. The advantage of a thesaurus or scheme, is that their elements have a restricted, well-defined and stable meaning, which addresses the most important problem with natural language. At the same time, this is also the most important drawback: new developments are covered conservatively, which makes it difficult to delineate such developments. Also, interdisciplinary research for which no specific terms in the vocabulary are (yet) available, may result in overly broad delineations because there is no appropriate, specific descriptor available[25]. An interesting hybrid between controlled and uncontrolled vocabulary are the Keywords Plus in the WOS, which are natural language terms from cited publications extracted in a controlled fashion (Garfield and Sher, 1993).

References

The last alternative we mention is the use of a core set of specially selected, important publications. An advantage is that a set of 'core' publications is usually not difficult to get: we could take highly cited publications in some area, or ask field experts. Using this core set, we can then either include publications which are bibliographically coupled (see **page 31**) with the publications in the core set, or include publications that cite the core set. An important advantage of this alternative is that we can make use of existing reference structures, which is up to date and independent from journals or disciplines. Moreover, we can try to reduce the amount of subjectiveness introduced by experts by using these structures. And if we choose to use expert knowledge, then we do not have the problems related to the vocabulary problem, as references are unambiguous. However, although references are unambiguous in the sense that they always refer to one publication, the *reason* for including a reference *is* ambiguous. As a result, we cannot trust that a single citation is sufficient to consider a citing publication part of a research area or field. Instead, at least a number of such citations are required, but it is a challenge to establish a minimum number of citations required. One approach may be to include publications in iterations and in each iteration require a higher number of citations in order to be included, as for example done in López-Illescas et al. (2009).

---

[25] Although we can use this to our advantage as well when trying to locate new developments, as we show in two papers in Part II of this thesis

1.4.5 Actors and prestige

People make science: researchers, educators, students, and lab assistants are all important to the craft and the institutions which harbour these people are at the hart of the system. These institutions have the responsibility to to organise and facilitate research, educate students, and not only act as "custodians of (new) knowledge, but also take responsibility for its dissemination" (Kauffman, 2009, p.239). In this large social construct, a central notion is that of 'academic prestige'. Individual publications can contribute to the prestige of actors[26] when these publications are cited often, based on the view that citations are judgements or votes of peers. Since citations are so important, it is an interesting question what gets a publication cited. In a study by Haslam et al. (2008), it was demonstrated that in the field of 'social and personal psychology', an eminent first author or a more senior later author, a prestigious journal, the length of an article (longer articles succeed better on average), and the inclusion of more recent references, were all good predictors of impact. On the other hand, having authors of both genders, of different nationalities, or from prestigious institutions, turned out to be no guarantee for success. Apparently, it is important for publications to be current (refer to recent publications) as well as recognisable (feature authors with a known track record[27]). Of course, this study only took into account specific features of a publication, and did not include quality of content in the analysis, a factor that will play an important role when peers judge a publication.

1.4.6 Organisation of science

Early in the history of modern science, institutions were organised along geographical and political lines, like the Royal Society in Great Britain and the Académie des Sciences in France. However, over time this expanded and changed, and science became primarily organised according to specialisation, such as physics, chemistry, psychology, or medicine. As we noted above in **section 1.3.1**, this kind of specialisation is a logical consequence of the scientific discourse, which clusters publications into groups with the same (kind of) topic. The result is a system that exhibits a fractal-like growth in which specialisation is a feature of its growth (Van Raan, 2000, p.359).

The structure of specialised disciplines, fields, and sub-fields, is what most people understand as the 'disciplinary' structure of science. However, this structure is also

---

[26] Actors is a general term for the people, institutions, and countries performing research and producing scientific publications.

[27] One of the more intriguing aspects of science (or anything where credit can be earned, for that matter) is that highly cited people are more likely to be cited even more. The phenomenon of "citations beget citations", which was first noted a long time ago by Merton (1968), is also called the "Matthew effect", after a quote in the gospel of St. Matthew: "*For unto every one that hath shall be given, and he shall have abundance: from him that hath not shall be taken away even that which he hath*" (Merton, 1968, p.159).

(negatively) associated with a certain conservatism, and is even considered as 'old fashioned'. The work of Gibbons et al. (1994) calls the familiar knowledge production within the boundaries of the disciplinary structure *Mode 1*, contrasting it with a 'new' *Mode 2*, which would operate "within a context of application in that problems are not set within a disciplinary framework [but are] transdisciplinary rather than mono- or multidisciplinary [and] carried out in non-hierarchical, heterogeneously organised forms which are essentially transient". We will explain the difference between multidisciplinary and transdisciplinary in the next section. One of the reasons for *Mode 2* is the massification of education in many countries, in which more competent (potential) researchers are educated than the disciplinary system can absorb. As a result, new sites were created outside the established institutes, like the Dutch TNO[28], its German counterpart the Fraunhofer Society, the American company Gartner[29], or the RAND[30] corporation. These institutions have a focus on applied research, but still take part in the production of knowledge and also compete with the 'traditional' disciplines on the global knowledge market.

From a policy perspective, the creation of multidisciplinary research teams seems appealing, as it could give a 'broader perspective' on problems and prevent lopsided solutions. Of course, this is not a silver bullet, and it requires careful attention and planning to make this approach succeed, as nicely formulated by Persson (1999, p.325): "A building consists of a lot of bricks, but a pile of bricks is not necessarily a building". However, since disciplines serve as the 'clearing house' for academic status and prestige, multidisciplinary research may find prestige more difficult to attain than monodisciplinary approaches. This appears to be the general feeling among academics, as investigated De Boer et al. (2006), and one that appears to be supported by findings from Carayol and Thi (2005), who found "evidence of no or negative incentives provided by the academic reward systems toward performing interdisciplinary research". However, Rinia et al. (2001a) found no (general) evidence of a negative bias to interdisciplinary research in peer-review assessments of physics, nor in assessments based on bibliometric indicators. Even more, Larivière and Gingras (2010) show that the amount of interdisciplinary links in a publication (often an indicator for multidisciplinary approaches) may influence impact, but that the type of influence may be either positive or negative, depending on the discipline in which the publication is published.

We regard the identification of two different modes of knowledge production as useful observation, but with a strong focuses on top-down co-ordinated, integrated

---

[28] TNO stands for 'Toegepast-natuurwetenschappelijk onderzoek', in English 'Netherlands Organisation for applied scientific research'.

[29] Gartner is an information technology research and advisory company, which was established in 1979 and features about 1,200 people in R&D according to its own website (`http://www.gartner.com/technology/about.jsp`).

[30] RAND, which stands for 'Research ANd Development', is a nonprofit research organisation, originally for the armed forces of the US government, but now also for many other governments as well as organisations and companies in many non-defence matters; according to its own website, the organisation has about 1,700 employees (`http://www.rand.org/about/people.html`).

approaches of problem-driven research, and disregarding “bottom-up processes originating from within disciplines” (Rinia, 2007, p.161). Such bottom-up processes exist, as De Mey (1982, p.175) notes, because researchers who are part of a specific community or discipline “strive to adhere more and more to the socially negotiated standard representations, while on the other hand they preserve individuality by the selection of a particular detail”. So, on the one hand a particular community agrees on certain formats and topics, while on the other hand individual members are required to add something new. One way to be novel is to import knowledge from another context and apply it locally, while another is to import problems from another domain and apply local knowledge (**section 1.5.2**). Such initiatives arise ‘spontaneously’, without top-down direction.

1.4.7 Levels of disciplinary integration

In the description of *Mode 2* research above, we note that a difference is made between different levels of integration in disciplinary cooperation. Following Van den Besselaar and Heimeriks (2001), we distinguish three different levels. The basic level is where a mix of people from different disciplines is working together, perhaps on a single project or topic, and concepts or knowledge from the contributing disciplines are used without integrating them. This level of cooperation is called multidisciplinary. In another level, a group still maintains strong links to the original disciplines, but has also started to share concepts and knowledge. This level is called interdisciplinary research, where there is an active interaction between the different disciplines. In a third level concepts and knowledge, though perhaps originating in contributing disciplines, have become integrated, up to a point that the research is more or less self-sustained and newly developed, shared concepts and knowledge has become the basis for further research. This final level is called transdisciplinary, as it ‘transcends’ the research in the originating disciplines.

1.4.8 On structures and context

Hitherto, we have used the term ‘structure’ mostly in the sense of a social organisation of science: the ‘disciplinary structure’. However, we can also think of other structures, such as ‘cognitive’ structures of themes and topics. or political structures of (temporary) alliances that want to achieve certain goals. In this thesis we will use ‘structure’ mostly in this wider sense, akin to what Merton formulated in his Foreword to Garfield and Merton (1979) as “the social, cultural, and cognitive structures latent in the practise of science” (*ibid* p.*v*). Moreover, we make a direct connection to scientific publications, and we define a *structure* as ‘a set of possibly overlapping groups of (elements from) publications’. These groups are implicitly understood as being composed of related elements, possibly a representation of some abstract object. For example, a structure can be a set of topics that make up a research area, a set of publications which share certain citations, or a network of authors that co-authored a set of publications. There is no unique structure for a set of publications, because a particular grouping[31] depends on both the element that is used to create groups, as well as the method that is used to create the groups (see **Section 1.6.3**).

When analysing structures in science, it is a challenge to make sense of why and how elements are grouped in a structure, and to assess whether groups make sense in the *context* of the research described by the publications. For example a group which represents research on oil refinery makes perfect sense in the context of chemistry or research into sustainable energy, but would be harder to place in the context of research into the behaviour of macaques. Filtering out elements from unrelated publications is sometimes tedious or difficult, especially when there are many publications involved. However, since unintentionally mixing up publications from different contexts can have an effect on the analysis and the calculation of indicators, it is usually important to invest some effort in their removal. Visualising the structure (**Section 1.6**) can be helpful, because groups of unrelated papers may appear as structures at the periphery of (for instance) a map.

1.4.9 Summary

Scientific publications are the result of the scientific discourse. There are several types of collections of publications, the broadest of which are the general bibliographic databases. An important general bibliographic database is the WOS, which is the basis for most of the work presented in Part II of this thesis. The creation of indicators for impact and output, an important part of science analysis. However, all analyses start by collecting relevant papers, which is called delineation. Such delineations can be based on journals, affiliations, natural language, controlled vocabulary, references and citations, or a combination of these. The notion of prestige is important and is one of the factors driving researchers to publish publications, as quality of work contributes to prestige. The scientific discourse is one of the reasons why science is organised in specialised disciplines and although contributions can come from outside established scientific institutions, these disciplines remain important ‘clearing houses’ for quality. Interdisciplinary work is associated with broad scope and innovative solutions. However, the academic appreciation for such initiatives remains uncertain and interdisciplinary approaches require careful monitoring. Next to disciplinary structures, we also have more general structures, which we abstractly define as a set of possibly overlapping groups of (elements from) publications.

1.5 Change and developments in science

In **Section 1.3** we noted that scientific knowledge is a collection of facts, methods and tools that is constantly changing: some elements become obsolete, others are

---

[31] We call a particular set of groups a ‘grouping’ or ‘constellation’ in this thesis.

adapted, and some are newly developed. In this section we give a brief overview on change and developments in science, with a strong focus on *new* developments and not so much on the ageing and obsolescence of research[32]. Below, when we use the word 'development', we mean the progression of a research context: the scientific discourse on certain, distinctive topics and the social and cognitive structures associated with this discourse. Specifically, a 'new development' is a 'change' in an established discourse, on new (combinations of) topics, which may require the creation of new structures and thus result in a new research context.

1.5.1 Phases of development

Global scientific output still steadily expands. For a long time, this appeared to go at an exponential rate, (De Solla Price, 1965; O'May, 1966 and Goffman, 1971), although recently Larivière et al. (2008) showed that this assumption may be too general. At lower levels of aggregation things are more turbulent: some research topics become obsolete and others come into vogue; sometimes whole research areas dissolve, while other ones get established (Van Raan, 2000). With respect to developments in science, one of the most cited works is Kuhn (1970). Among other things, the 'Kuhnian model' distinguishes two general phases: an adaptive phase ('normal science') and a disruptive phase ('paradigm shift'). The work is widely commented on. Crane (1980) for instance, studied the development of high energy physics in the context of the Kuhnian model, and she noted that although the basic elements of the model appeared to exist (in high energy physics), they may work differently than originally proposed. For example, paradigms and values are shared by the whole research area and not only by a small community. Another comment was made by Van Raan (2000) who investigated the development of the global output of scientific publications in the twentieth century, and concluded that there appeared no evidence for large paradigm shifts.

Whether or not the nature of a development is adaptive or disruptive, having different phases is something that is common to most developments involving time or quantities. For example, in the lifetime of a product or technology, Solomon et al. (2000) discerns six phases: 1) introduction; 2) growth; 3) maturation; 4) decline; 5) phase-out; and 6) obsolescence. In scientific developments, such phases have been noted as well, for example in mathematics by O'May (1966). Focusing on introduction, growth, and maturation, Tabah (1999, pp.274-5) proposed the following phases in the development of a research area:

1. birth of a notion;
2. spread and initial uptake;
3. acceptance and subsequent fragmentation into subtopics; and

---

[32] An interested reader is referred to for example Egghe and Rousseau (2000) for a more extensive coverage of the topic.

4. codification, institutionalisation, and reorganisation.

Here, ‘growth’ is divided in two: ‘spread’ and ‘acceptance’. The difference is that after spreading, a topic has to be established and shown resilient to initial (critical) examination for subsequent growth in the acceptance phase to built on. In the initial phases of a development, and especially in the case of a development involving different disciplines, the common ground will be in a state of flux. Once a new development has moved beyond this phase, it will start to show a more ‘coherent’ or integrated structure where journals, references, and jargon have a use and meaning that is mostly agreed upon.

1.5.2 Cause and establishment

Any discourse on a topic has a common ground on which it is based, with elements that (in the sense of solutions to problems) are also called exemplars in the Kuhnian model (see also **Section 1.3.4**). Morris (2005, p.1252-3) notes six different types of exemplars in scientific developments:

1. discovery or introduction of a new concept;
2. borrowed knowledge;
3. empirical knowledge;
4. extension of concepts;
5. application outside the area;
6. consolidation, such as a review.

Any combination of these exemplars can also represent the *cause* of a development. However, larger objectives (either internal or external to science) can cause developments as well. An example of an external objective is the societal desire for cleaner energy, while the goal of a ‘theory of everything’ is an example of an internal objective. Developments that are caused externally may have their own character: in the research for one of the publications in this thesis, we noted that the birth of ‘sustainability science’ was not announced in one, but in a number of publications, including (and probably not limited to) Lee (1993), Dodds (1997), Kates et al. (2001), Clark (2001), Clarke (2002) and Clark and Dickson (2003).

To get established, a new development has to ‘prove’ itself by showing prospect. The peer-review system of science and the importance of attaining prestige, are two factors which contribute to a certain cautiousness in science, which is important to keep the growth of output in check while maintaining a certain level of quality. At the same time, this can limit the uptake of new developments, because when the first initiatives are published they may not appear to have enough potential to be worth looking into. Additionally, social, political, and cultural facets of science can create barriers which have nothing to do with peer review, quality, prestige, or potential (Ruiz-Baños et al., 1999).

Knowledge can be borrowed from a neighbouring, related research area, or from a completely different field or discipline, for instance due migration of scholars from one field to another (Urata, 1990). If results of such developments are also taken up in the original field(s), then this is called mutually enabling or 'converging research' (Roco and Bainbridge, 2003 and Nordmann, 2004). Convergence can be regarded as different common grounds (see **page 29**) which start to overlap and (when given time and the research is successful) could even merge. One of the prime examples of converging research is between computer science and nano-science. New developments that involve different disciplines may go through different levels of cooperation (**Section 1.4.7**): start out multidisciplinary, going through a phase of interdisciplinary research, while finishing at a level of self-sustainability and transdisciplinarity.

New developments involving authors from other fields or disciplines can face additional challenges. Palmer (1999) notes that even experienced scientists feel challenged by the task of acquiring knowledge from outside their realm of expertise. Also, when a development takes place by translating knowledge from one field to another, such a translation requires considerable effort by the researchers responsible for the translation. Still, such translations are at the root of many new developments, either as the result of creative 'probing' by researchers, or of multidisciplinary efforts related to tasks, processes, products or use (Porter and Chubin, 1985).

1.5.3 Searching new developments

Modelling growth and establishment of new developments in science (O'May, 1966; Goffman, 1971 and Morris, 2005) results in general characteristics of that growth and establishment. Such characteristics can be used to *search* for these developments. One important characteristic of a new development is its fast growth, which can for example be found by fitting growth curves (Egghe and Ravichandra Rao, 1992). To fit curves one needs a large enough number of publications, which may not always be the case. So alternatively, fast growth can be quantified and measured by an indicator such as the time it takes to double the number of papers. Another alternative is to search for changes in structure. For example, an increasing number of citations from one established field to another, or number co-cited authors from different fields (Small, 2006) could indicate such a change. Another example is the creation of new journals, to accommodate a newly establish discourse, although this usually implies that the new development has reached a phase which justified the investment in a journal. Also, publications or (publications related to) specific topics that are cited a lot in a short period of time ('hot topics') can also indicate new developments. The last alternative is to use background knowledge, either by directly interviewing experts, or indirectly, by reading editorials in reputable multidisciplinary journals such as Nature and Science.

However, searching has a number of additional challenges. For one, not all new developments continue after birth or initial uptake. So, we run the risk of finding 'false positives': new developments which will decline after an initial (quick) uptake and spread. Another challenge is to quantify phases in terms of output. For instance, we cannot easily say when a topic has a 'large enough' number of publications to consider it beyond an initial uptake. Also, if there would be such a number, it would probably differ from field to field. A quick scan of the literature describing 'emerged' (newly developed) fields shows this. For example, the study of Hinze (1994) on the emergence of bioelectronics includes about 250 publications, while the study of Berg and Wagner-Döbler (1996) on the research on fuzzy sets includes only about 100 publications. Added to that, early in the spread of a topic, these numbers can be even smaller. For instance, in the first year the bioelectronics study by Hinze (1994), there are only ten publications. Finally, there is another, more fundamental challenge: how to pick those publications which (taken together) show such a curve. At the level of major fields for example, developments on specific topics are lost in the large numbers of publications that appear each year. In two of our contributions in Part II, we have solved this by focusing on finding converging research. Additionally, we have used the distribution of a growth indicator to differentiate between high and low growth.

1.5.4 Summary

In general, science still expands fast. At lower levels of aggregation developments take place in phases of growth and decline. A new development can be started by a new discovery, the introduction of new concepts, knowledge, or as a result of larger objectives. After having been established, a new development is spread, accepted, and codified. Acceptance may show increasing levels of integration and self-sustainability. A challenge is to quantify the size of a development that is worth noting. Science is cautious in accepting new developments, and developments from different research contexts may face additional challenges. Still, new developments can be found using growth characteristics, locating structural changes, or by using background knowledge.

1.6 Visualisation and maps

The saying goes that "a picture is worth a thousand words". Not only can we remember a picture much easier than a table filled with words and numbers, our ability to quickly see patterns in images can be a great aid in the analysis of complex data. Classic examples in Tufte (1983) show how visualisations can quickly reveal patterns which would otherwise require more effort to discern. When we analyse science publications, we have to deal with a large body of knowledge, mostly encoded in text, which cannot be viewed in its entirety as it is too large and complicated. As White and McCain (1997) note in their seminal paper, this is exactly why visualisation is such an important tool to analyse science, because it can give an overview of such a body by aggregating over details, but preserving the most relevant relations[33].

1.6.1 Content and references

Besides giving excellent examples, Tufte (1983) also stresses the importance of *straightforward* visualisations that avoid unnecessary decorations. This advice is partly of an aesthetic nature, but it also serves a function which we can understand in terms of what we write in **Section 1.3** on meaning and context. A visualisation of scientific publications contains elements (points, lines, words, colours) which are interpreted as signs that refer to abstract objects such as research topics or concepts. When two such elements are plotted close together, they tend to be interpreted together and thus influence their mutual *as well as their own* interpretation. To give a classic example of how unnecessary decorations can confuse, we show in **Figure 2** a chart in which we (fictitiously) plot people against a room number in a hotel. The line mistakenly creates the impression that the points are related, yet the assignment of rooms is of course (most of the time) completely independent of the subsequent persons occupying it. Still, although the use of unintentional (or unnecessary) context should be avoided, the *intentional* use of context is what gives many (if not most) visualisations their actual power.

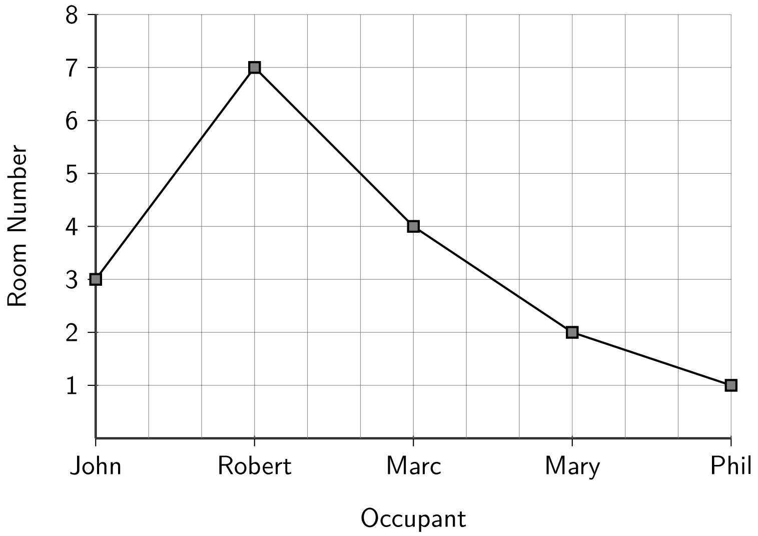


Figure 2 An example of a visualisation in which context is used improperly, as the line between the points creates an unintended impression that the points are related. This is a fictional example that plots the names of people who have been assigned a room number.

---

[33] This is similar, though a bit more specific than of Williams et al. (1995), who call it a means to help create a "mental image of a domain space" (*ibid* p.163).

To be useful, elements in a visualisation should refer to objects known to a person that has to interpret the image. If this is not (or only partly) the case, then the visualisation will fail to transfer its message. Young children can for instance be baffled by visualisations which pose no problem for adults, because they miss the connection between pictured elements and objects (DeLoache and Marzolf, 1992). However, as White and McCain (1997) note in their gentle critique, this may also happen if not enough attention is given to the user-friendliness of the visualisation, for example by "depending too much on unlabeled points in space, mysterious shapes, number codes, color codes, or barebones abbreviations [...]" (*ibid*, p.145). Unfortunately, this critique is still relevant today, even (or especially?) for many academic contributions that propose or employ visualisations to analyse complex knowledge domains.

1.6.2 Maps and networks

Visualisations need not to be complex to be effective. Simply labelling parts of a plot can already highlight a connection between growth and research area, as shown by Noyons and van Raan (1998, p.79). Most visualisation are more complex though, by which we mean that the amount of information compressed into a picture is larger than in (for instance) a labelled plot. The two most used types of visualisation in quantitative analyses of science, maps and networks[34], are good examples of more complex visualisations. Maps are images in two or three dimensions[35] of elements which represent (groups of) items, in which the distance between two elements on a map expresses the (dis-) similarity between the items that the elements represent. Science maps can represent items such as actors, noun phrases, or references. Quantitative maps are created algorithmically, based on a square matrix with numerical data, where the rows and columns are the items displayed in the map. We distinguish such maps from qualitative maps which are created manually or semi-automatically, based on some internal cognitive structure (see **Section 1.3.3**), representing for example the construction of a research field, or the interpretation of a piece of text. We will get back to qualitative maps in **Section 1.6.4**

Maps and networks share important features: both can represent abstract items, both can be created algorithmically. However, there is an important *conceptual* difference between the two: the role of the relation between elements is *central* to networks, while much less so to maps. Also, to create maps we are usually restricted to distance-preserving layout algorithms (such as MDS[36]), whereas networks can employ a larger

---

[34] When we use 'networks' in this section, we mean the visualisation of such networks and not the network data.

[35] However, three dimensions are currently not used as much, mostly because the current plotting devices are not stereoscopic and are thus not able to plot in "real" 3D—only recently have such devices started to become widely available.

[36] MDS stands for 'Multi Dimensional Scaling', an algorithm which projects a distance matrix into two dimension, in such a way that the distance between the vectors in the matrix is preserved as much as possible. (Kruskal, 1964)

range of both distance-preserving (e.g. MDS, Kamada and Kawai (1989), and more recently Van Eck and Waltman (2007)) as well as algorithms that do not preserve distance (for instance laying out a network in a circle, e.g. Klavans and Boyack (2009)). Additionally, many social networks represent unweighted relations[37], whereas maps usually represent weighted data. We have seen an increase in interest in (social) networks in bibliometric analyses over the last ten years (Otte and Rousseau, 2002), both as a tool for visualisation and clustering. However, since the work we present in this thesis uses mostly maps, we will in this general overview therefore focus on this type of visualisation.

1.6.3 Creation of maps

As we noted above, maps are representations of square matrices of (dis-) similarities between items that either refer directly to scientific objects or are proxies for these objects (see **section 1.3.1** on **page 25**). For example, phrases can represent cognitive aspects such as topics or concepts (Noyons, 1999), while actors can represent social aspects. Also, items are not restricted to a single type, as explored by Braam et al. (1991a) and more recently by Janssens et al. (2009), although this adds to the complexity of the visualisation. Whatever type is used, only those items should be selected that best represent objects in the research represented by the scientific publications. This is a challenging subject, similar to the challenges we face when delineating a set of publications. Preferably, a selection should be as objective and (hence) require as little manual selection as possible, but this is still an active research topic (Van Eck et al., 2010; Ozmutlu, 2006 and Tseng et al., 2009).

Items can be displayed directly on a map or can first be assembled (clustered) into groups (clusters) of related items, after which these clusters can be plotted instead. The advantage of grouping items is that the display becomes less cluttered and better to comprehend. A disadvantage is that information is lost, because the structure inside a group is not displayed, only the aggregated distance of items in a group relative to those in other groups. There are many algorithms to cluster items. An extensive overview by Jain et al. (1999) reviews algorithms like hierarchical clustering, nearest neighbour, fuzzy clustering, neural networks, and genetic algorithms, and an overview of clustering algorithms in the context of domain visualisation is given by Börner et al. (2003). Currently, a popular choice are algorithms developed in the context of network analysis (Girvan and Newman, 2002 and Waltman et al., 2010). In our work, we have mostly deployed hierarchical clustering, because of fast implementations and a well-established use in science analysis. Added to that, the resulting dendrogram (tree of items) is easily assessed and can be helpful in additional (meta) analysis[38]. A challenge with hierarchical clustering however, is to find the appropriate number

---

[37] Unweighted relations either exist or they do not and a matrix representing such relations contains only zeroes and ones.

[38] In the context of clustering, a meta analysis is an additional clustering of the clusters used in the map.

of clusters. We have employed a number of approaches, such as calculating statistics promoted by Milligan and Cooper (1985) and Tibshirani et al. (2001), or repeated sampling and clustering, as developed by Ben-Hur et al. (2002). Unfortunately, we have not yet found a definite solution, and we are still actively researching the topic (Van Eck et al., 2010). Clustering algorithms such as unsupervised neural networks employed in for example Polanco et al. (2001), do not have this problem, at the cost more complexity of both algorithm and the subsequent result. A clustering algorithm may divide items into either crisp or fuzzy groups, which means that items are either assigned to a single group or to (possibly in varying degree) different groups. As a result, publications from which items are extracted are (indirectly) clustered based on their relation with the items. If this relation is $1 : N$, for instance when we use NPs (**page 26**), then publications may be associated with NPs in more than one group; thus, even if NPs are crisply divided into groups, the publications in which they occur are not. However, if the relation between a publication and an item is $1 : 1$, for instance when we use journals, then publications are grouped identically to these items.

1.6.4 Qualitative maps

Not all visualisations are created automatically, nor are they restricted to quantitative data. For example, conceptual structures (Sowa, 1983) can be used to represent knowledge, conceptual graphs (Hoede, 1994) to represent meaning (semantics) of a natural language phrase, and conceptual maps (Huff, 1990) to represent the (interpretation of) a piece of text. In these maps elements (nodes) are placed in the map by a user, instead of by an algorithm, and the reason a user places a node at a particular place may involve more than the 'relatedness' between nodes. For example, a user may find a given concept central to the picture, thus placing it at the centre, or a user may structure a map according to different areas of interest. Qualitative maps are interesting because they are a reflection of how a user sees (and values) a structure, understands the meaning of a phrase, or interprets a text.

The ability to structure information in a qualitative map is also useful in education, mostly as an alternative for 'linear' or 'root' learning (Novak, 1998). One of these tools is the 'woven stories' concept (Gerdt et al., 2001), which is a network of stories by (potentially) different authors. The stories are maps similar to those in Huff (1990), but with at least a number of differences: first, the nodes in woven stories usually contain more text; also, they feature a variable and optional pragmatic (see **section 1.3.2**) aspect, such as 'Introduction' or 'Method', as opposed to a non-optional aspect, selected from a limited number of types such as 'Main entity' or 'Related concept' Huff (1990, p.40); and finally, the nodes can have a specific ordering (Introduction, Method, Results, etc.).

1.6.5 Use of maps in science analyses

White and McCain (1997) note that there are two ways in which a visualisation may be used in an analysis of a knowledge domain: 1) domain analysis; and 2) information retrieval (IR). Also, they list a number of specific tasks (*ibid* p.107) which can be perform in such analyses:

- locate 'hot-spots', like highly influential actors, publications, topics;
- identify relations between items;
- discover communities or groups of related items;
- identify trends.

When analysing a complex knowledge domain, both analysis and information retrieval is useful. Also, any in-depth analysis probably uses most tasks in concert. Below, we list three types of analyses of science that commonly employ visualisations.

Scientific domain analysis

In scientific domain analysis the structure displayed by a map is used together with the ability to retrieve information about items represented in the map (Noyons, 1999). This is one of the areas where some of our work in Part II has found much application and has also been extended in the academic literature, like in Moya-Anegón et al. (2004) and Mothe et al. (2006). Although domain analysis makes use of a map to retrieve information, this has a different goal than when using a map display only for information retrieval, as for instance in Lin (1997): instead of trying to locate relevant literature we want to establish what *science* is represented by a specific item or group of items. So, instead of (only) titles of relevant documents, a user may query for the most prevalent journals, the best cited authors, or the noun phrases which were clustered in an item. It is also different from the analysis featured in Stockwell et al. (2009) who also use cognitive mapping to analyse (scientific) publications, but who are primarily interested in exploring the content of the publications, as opposed to exploring the science represented by the mapped content.

Scientometric assessment

Another application of maps in the analysis of science is to visualise bibliometric indicators on top of items in a map. Such an application is also useful in a scientometric assessment, as explored by Noyons et al. (1999). The structure of a map is used to differentiate indicators over elements in a map, thus providing a more detailed view of the performance of an actor. For example, an actor can perform on average in some field, but perform excellent in a limited number of topics. A map can highlight this differentiation in performance if these topics are represented in a map (preferably as separate elements). In a sense, this can improve the 'fairness' of an assessment.

Changes in scientific structure

A third application of maps is in the detection of changes in scientific structure. In general, we find two ways of presenting such changes. One way is to show the development of the size of elements in a map or the strength of the relations between elements using plots and charts, as done in for example Braam et al. (1991b), Chavalarias and Cointet (2008), and more recently by Takeda and Kajikawa (2009). Another way is take snapshots of maps at different moments in time and to visually inspect the changes between snapshots (Chen et al., 2002). Alternatively, the changes in position can become part of the visualisation (Noyons and van Raan, 1998). A combination of both is used by Shibata et al. (2008). Of course, animating changes can be highly informative, but we find this of limited application in the literature as the current format of scientific publications still precludes including animations.

1.6.6 Use and validity of maps

We can see a map both as a tool and as a model of a complex body of (scientific) knowledge. As such, a map has two types of validity: one is the extent to which a map is usable for the purpose it was created for (utility), next to being a 'true' representation of the depicted research (structural, see also the introduction of Noyons (1999)). To judge the utility aspect, we need to assess whether a map can be used for its particular goal, such as differentiating research performance for actors in different topics, while to judge the structural aspect of validity, we need to assess whether a map is displaying the research area properly. Utility aspects are usually not too difficult to judge. However, to judge the structural validity of a map is more difficult, as this requires another structural representation to compare with.

In IR, the performance of a model is usually assessed computationally, by using a specifically prepared 'test set' containing pre-classified documents or terms that are used to calculate precision and recall. In computer science, programs are tested using test programs, some of which validate the output of small parts of a program ('unit tests'), while other programs validate the flow of inputs and outputs ('scenario tests'). This is not applicable to maps, because sets of examples that are marked with a 'correct' clustering are difficult (if not impossible) to create. What we can do automatically is to test the stability of maps by re-sampling sets of data used to create the map and test whether subsequent re-samplings result in approximately the same map (Ben-Hur et al., 2002). Such a test would be able to express whether a map is the result of some random constellation of (items from) publications.

However, even after we establish that a map is a non-random constellation, we still have to make sure that the visualisation can be regarded as a model of some non-random collection of publications. For most scientific fields, this implies that we need the help of experts, who compare the map with their internal representation (**Section 1.3.3**) of the mapped research. For non-experts, such a judgement by experts become 'a seal of approval'. In a sense, these experts become secondary, trusted indices to the knowledge represented by a map. However, just as the indexing terms investigated by Mai (2001) (which we mention on **page 25**) are the result of an interpretive process, so is the judgement of an expert with respect to a map. Added to that, a map can neither be true nor false: at most, its configuration of symbolic references can be seen as a good or bad 'fit' to the world view of an expert. Moreover, if regarded as a model, then a map shares the limitations of all abstract representations. First of all, it is by design an *approximation* of a highly complex structure which (again by design) includes only a fraction of the available information. Next to that, the elements in a map represent (items in) scientific publications, which are assumed to refer to objects such as research topics or concepts. However, a person who judges a map may not recognise the items as references to these objects, or may consider the represented objects not part of the field. The same applies if items are (signs of) proxies (instead of direct signs) and a person does not recognise the chosen proxy to refer indirectly to some type of object. So, even if one expert would consider a map a structurally valid representation, there is no guarantee that another will do so as well. In a similar vein, Healey et al. (1986) noted already some time ago a 'disjunction' between map makers and their clients, as well as an apparently fundamental 'validation paradox': "[i]f the results of the work are counterintuitive to experts they are considered invalid; if the same as their usual intuitions, they are considered valid but uninteresting—they refer only that which is already known." (*ibid* p.247) Of course, this paradox only becomes a problem if expert opinion is the sole basis of confidence a map.

In this respect, it is informative to look at the difference between knowledge acquired from maps and knowledge acquired from navigating a mapped area (without map), as investigated in Thorndyke and Hayes-Roth (1982). They show that people acquire 'survey knowledge' (a 'bird's eye view') by using a map, while people acquire *procedural knowledge* (how to traverse an environment) by navigating the mapped area instead. The survey knowledge is superior when asked to quickly explain the (spatial) relations between objects. The difference is nicely illustrated by the difficulty people experience when requested to draw a map of a familiar environment (*ibid* p.585). The reason why people find this difficult is that it requires a change of perspective: from being 'inside' the environment to being 'above' it. Now, if we assume that the analogy between maps of science and their geographical counterparts holds then this result also provides a clue why experts sometimes find it difficult to assess maps: they acquired knowledge by experience, which is thus of a procedural nature and may not immediately translate into the survey knowledge required to quickly assess a map. Next to that, a map may fail to show landmarks that are also prominent in the procedural knowledge of an expert, either because they are missing or have been obscured because they are in a group of related items.

Finally, most users of science maps already have some internal notion of the structure of the research being modelled—they are not clueless users. So, apart from the *external* validation by experts, users also validate *internally*, by looking whether it confirms to *their* expectations.

1.6.7 Summary

Visualisations can create comprehensive overviews of complex bodies of knowledge and can thus be very useful when analysing sets of scientific publications. Although there are many types of visualisations, maps and networks are most used in quantitative analyses of science. Maps are representations of (groups of) items, in which the items are placed in such a way that their distance reflects their (dis-) similarity as good as possible. Grouping these items may improve the comprehension of a map at the cost of loosing some information. Qualitative maps can be used to structure qualitative information and is also helpful in education. A map can be used to identify influential items, relations, communities (groups of related items), and also trends. These tasks can help to perform scientific domain analysis, present detailed scientometric results, as well as find changes in scientific orientation over time. To judge the validity of a map, a difference must be made between the structural aspect of validity and the utility aspect. To judge the structural validity, users may rely on their own knowledge of a field, or trust the judgement of a field expert, although the procedural knowledge of experts may not immediately translate to a map. The utility aspect is usually much easier to assess.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*(4), 249–277.

Barber, J. A. and Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine, 19*(23), 3219–3236.

Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17. Singapore.

Berg, J. and Wagner-Döbler, R. (1996). A multidimensional analysis of scientific dynamics. *Scientometrics, 35*(3), 321–346.

Börner, K., Chen, C. and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991b). Mapping of Science by Combined Co-Citation and Word Analysis II: Dynamic Aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351–360.

Carayol, N. and Thi, T. U. N. (2005). Why do academic scientists engage in interdisciplinary research?. *Research Evaluation, 14*(1), 70.

Chalmers, A. F. (1999). *What Is This Thing Called Science?*. Buckingham: Open University Press.

Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine, 15*(13), 1447–1458.

Chavalarias, D. and Cointet, J. (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. *Scientometrics, 75*, 37–50.

Chen, C. M., Cribbin, T., Macredie, R. and Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal of the American Society for Information Science and Technology, 53*(8), 678–689.

Clark, W. C. (2001). Research systems for a transition toward sustainability. In Steffen, W., Jäger, J., Carson, D. J. and Bradshaw, C., editors, *Challenges of a changing earth: Proceedings of the global change open science conference, Amsterdam, 2001*, pages 197–200. Berlin.

Clark, W. C. and Dickson, N. M. (2003). Sustainability science: The emerging research program. *Proceedings of the National Academy of Sciences of the United States of America, 100*(14), 8059–8061.

Clarke, T. (2002). Sustainable development: Wanted: scientists for sustainability. *Nature, 418*(6900), 812–814.

Crane, D. (1980). An exploratory study of kuhnian paradigms in theoretical high energy physics. *Social Studies of Science, 10*(1), 23–54.

Cronin, B. (2000). Semiotics and evaluative bibliometrics. *Journal of Documentation, 56*(4), 440–453.

De Boer, Y., De Gier, A., Verschuur, M. and De Wit, B. (2006). Building bridges: Researchers on their experiences with interdisciplinary research in the netherlands. Technical Report Den Haag: RMNO.

De Bruin, R. E. and Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics, 26*(1), 65–80.

De Mey, M. (1982). *The Cognitive Paradigm.* London: D. Reidel Publishing Company.

De Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149*(3683), 510–515.

DeLoache, J. S. and Marzolf, D. P. (1992). When a picture is not worth a thousand words: Young children's understanding of pictures and models. *Cognitive Development, 7*(3), 317–329.

Dodds, S. (1997). Towards a 'science of sustainability': Improving the way ecological economics understands human well-being. *Ecological Economics, 23*(2), 95–111.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Egghe, L. and Ravichandra Rao, I. K. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics, 25*, 5–46.

Egghe, L. and Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science, 51*(11), 1004–1017.

Egghe, L. and Rousseau, R. (2002). A proposal to define a core of a scientific subject: A definition using concentration and fuzzy sets. *Scientometrics, 54*(1), 51–62.

Figala, K. (2002). Newton's alchemy. In Cohen, I. B. and Smith, G. E., editors, *The Cambridge Companion to Newton*, pages 370–386. Cambridge: Cambridge University Press.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.

Garfield, E. and Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. New York: Wiley.

Garfield, E. and Sher, I. H. (1993). Keywords plus algorithmic derivative indexing. *Journal of the American Institute for Information Science, 44*(5), 298–299.

Gerdt, P., Kommers, P., Looi, C. K. and Sutinen, E. (2001). Cognitive technology 2001. In Beynon, M., Nehaniv, C. L. and Dautenhahn, K., editors, *LNAI 2117*, pages 233–247. Springer-Verlag, Berlin.

Gibbons, M., Limogenes, C., Nowotny, H., Schwartzman, S., Scott, P. and Trow, M. (1994). *The new production of knowledge.* Sage Publications.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821.

Goffman, W. (1971). A mathematical method for analyzing the growth of a scienfic discipline. *Journal of the Association for Computing Machinery, 18*(2), 1672–1673.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J. and Wilson, S. (2008). What makes an article influential? predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.

Healey, P., Rothman, H. and Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233–251.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology, 42*(2), 182 –192.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: the case of bioelectronics. *Scientometrics, 29*(3), 353–376.

Hirsch, J. E. (2007). Does the h index have predictive power?. *Proceedings of the National Academy of Sciences, 104*(49), 19193–19198.

Hirst, G. (1978). Discipline impact factors: A method for determining core journal lists. *Journal of the American Society for Information Science, 29*(4), 171–172.

Hoede, C. (1994). On the ontology of conceptual graphs. Memorandum 1239, Enschede: Faculty of Applied Mathematics, University of Twente.

Hofstadter, D. (1985). *Metamagical Themas: Questing For The Essence Of Mind And Pattern.* New York: Basic Books.

Huff, A. S. (1990). *Mapping strategic thought.* Chichester: John Wiley & Sons.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264–323.

Janssens, F., Zhang, L., De Moor, B. and Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management, 45*(6), 683–702.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters, 31*(1), 7–15.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Grübler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., Mooney, H., Moore, B., O'Riordan, T. and Svedin, U. (2001). Sustainability science. *Science, 292*(5517), 641–642.

Kauffman, J. (2009). Advancing sustainability science: report on the international conference on sustainability science (ICSS) 2009. *Sustainability Science, 4*(2), 233–242.

Klavans, R. and Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455–476.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115–129.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Landesman, C. (1997). *An Introduction to Epistemology.* Oxford, UK: Blackwell Publishers Ltd..

Larivière, V., Archambault, É. and Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology, 59*(2), 288–296.

Larivière, V. and Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126–131.

Lee, K. N. (1993). *Compass and Gyroscope: Integrating Science And Politics For The Environment.* Washington: Island Press.

Lewison, G. (1999). The definition and calibration of biomedical subfields. *Scientometrics, 46*(3), 529–537.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40–54.

López-Illescas, C., Noyons, E. C. M., Visser, M. S., Moya-Anegón, F. D. and Moed, H. F. (2009). Expansion of scientific journal categories using reference analysis: How can it be done and does it make a difference?. *Scientometrics, 79*(3), 473–490.

Mai, J. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation, 57*(5), 591–622.

Merton, R. K. (1968). The Matthew Effect in Science. *Science, 159*(3810), 56–63.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Moed, H., Bruin, R. D. and Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Moed, H. F. (2005a). *Citation Analysis in Research Evaluation*, volume 9 of *Information Science and Knowledge Management*. Dordrecht: Springer.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, cocitation and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, 56*(12), 1250–1273.

Mothe, J., Chrisment, C., Dkaki, T., Dousset, B. and Karouach, S. (2006). Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, Environment and Urban Systems, 30*, 460–484.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E. and Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics, 61*, 129–145.

Nordmann, A. (2004). Converging technologies — shaping the future of European societies. Technical Report EU HLEG Foresighting the New Technology Wave.

Novak, J. D. (1998). *Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations*. New Jersey: Lawrence Erlbaum Associates.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool*. PhD thesis, Leiden: Leiden University.

Noyons, E. C. M., Moed, H. F. and Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115–131.

Noyons, E. C. M. and van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

O'May, K. (1966). Quantitative growth of the mathematical literature. *Science, 154*, 1672–1673.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441–453.

Ozmutlu, S. (2006). Automatic new topic identification using multiple linear regression. *Information Processing & Management, 42*(4), 934–950.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science, 50*(3), 242–253.

Persson, G. (1999). Mistra's role and some lessons learned in the first five years. *Policy Sciences, 32*(4), 323–326.

Polanco, X., François, C. and Lamirel, J. C. (2001). Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. *Scientometrics, 51*(1), 267–292.

Porter, A. and Chubin, D. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3), 161–176.

Rinia, E. J. (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer*. PhD thesis, Leiden: Leiden University.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2001a). Citation delay in interdisciplinary knowledge exchange. *Scientometrics, 51*(1), 293–309.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics, 54*(3), 347–362.

Roco, M. C. and Bainbridge, W. S., editors (2003). *Converging Technologies for Improving Human Performance*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ruiz-Baños, R., Bailón-Moreno, R., Jiménez-Contreras, E. and Courtial, J. (1999). Structure and dynamics of scientific networks. part i: Fundamentals of the quantitative model of translation. *Scientometrics, 44*(2), 217–234.

Saeed, J. I. (2003). *Semantics*. Malden (MA): Blackwell.

Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*(11), 758–775.

Simar, L. (1992). Estimating efficiencies from frontier models with panel data: A comparison of parametric, non-parametric and semi-parametric methods with bootstrapping. *Journal of Productivity Analysis, 3*, 171–203.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*, 595–610.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327–340.

Solomon, R., Sandborn, P. A. and Pecht, M. G. (2000). Electronic part life cycle concepts and obsolescence forecasting. *IEEE Transactions on Components and Packaging Technologies, 23*(4), 707–717.

Sowa, J. F. (1983). *Conceptual structures: Information processing in mind and machinE*. Reading, Mass.: Addison-Wesley.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy, 25*(5), 701–721.

Stockwell, P., Colomb, R. M., Smith, A. E. and Wiles, J. (2009). Use of an automatic content analysis tool: A technique for seeing both local and global scope. *International Journal of Human-Computer Studies, 67*(5), 424–436.

Tabah, A. N. (1999). Literature dynamics: studies on growth, diffusion and epidemics. In Williams, M. E., editor, *Annual Review of Information Science and Technology*, pages 249–286. Information Today Inc.

Takeda, Y. and Kajikawa, Y. (2009). Optics: a bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics, 78*(3), 543–558.

Thorndyke, P. W. and Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology, 14*(4), 560–589.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.

Tijssen, R. J. W. (1992). *Cartography of Science: Scientometric Mapping with Multidimensional Scaling Techniques*. PhD thesis, Leiden: Leiden University.

Tseng, Y., Lin, Y., Lee, Y., Hung, W. and Lee, C. (2009). A comparison of methods for detecting hot topics. *Scientometrics, 81*(1), 73–90.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press.

Uhlenbeck, G. E. and Goudsmit, S. (1925). Ersetzung der hypothese vom unmechanischen zwang durch eine forderung bezüglich des inneren verhaltens jedes einzelnen elektrons. *Die Naturwissenschaften, 13*, 953–954.

Urata, H. (1990). Information flows among academic disciplines in Japan. *Scientometrics, 18*, 309–319.

Van den Besselaar, P. and Heimeriks, G. (2001). Disciplinary, multidisciplinary, interdisciplinary: Concepts and indicators. In *Paper the 8th ISSI Conference for Scientometrics and Informetrics, Sydney* .

Van Eck, N. J. and Waltman, L. (2007). VOS: a new method for visualizing similarities between objects. In Decker, R. and Lenz, H. J., editors, *Advances in Data Analysis*, pages 299–306. Berlin, Heidelberg: Springer Berlin Heidelberg.

Van Eck, N. J., Waltman, L., Noyons, E. C. M. and Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581–596.

Van Leeuwen, T. N. (2004). *Second generation bibliometric indicators – the improvement of existing and development of new bibliometric indicators for research and journal performance assessment procedures*. PhD thesis, Leiden: Leiden University.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.

Van Raan, A. F. J. (2000). On growth, aging and fractal differentiation of science. *Scientometrics, 47*(2), 347–362.

Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J. and Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. *Journal of Informetrics, 4*(3), 431–435.

Voutilainen, A. (1995). NPtool, a detector of english noun phrases. *cmp-lg/9502010*.

Waltman, L., van Eck, N. J. and Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S. and Van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47.

White, H. D. and McCain, K. W. (1997). Visualization of literatures. In Williams, M. E., editor, *Annual review of information science and technology (ARIST)*, pages 99–168. Information Today, Inc. for the American Society for Infomation Science.

Williams, J. G., Sochats, K. M. and Morse, E. (1995). Visualization. In Williams, M. E., editor, *Annual review of information science and technology (ARIST)*, pages 161–207. Information Today, Inc. for the American Society for Infomation Science.

Wouters, P. (1998). The signs of science. *Scientometrics, 41*, 225–241.

Zitt, M. and Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information processing & management, 42*(6), 1513–1531.

2

Bridge to Part II

2.1 Important topics in the general research background

In the previous chapter we have described the general background of our research. The following list reiterates some important topics and statements.

- Science is a complex phenomenon that plays a central role in modern society and has many facets.
- There is a fundamental gap between intangible aspects of science like knowledge and quality, and tangible aspects like publications, reports, websites and institutions.
- The scientific discourse is an elementary aspect of science, and scientific publications can be viewed as tangible representations of communication and knowledge, while citations (references) can be regarded as representing aspects such as importance and relevance.
- Publications are written and interpreted in the context of a ‘common ground’, which consists of a personal world view and shared background knowledge. The sign-triad is an abstraction to illustrate this, by showing that an interpretant links a tangible (encoding of a) sign with an intangible object.
- Common ground can be represented by the use of signs such as jargon, exemplars, and renowned actors. Such signs are usually particular to a research area or field.
- To conduct a useful scientific discourse, it is required to refer to common ground or to more specific parts of scientific knowledge. This is an important reason for both cognitive and social structures in science.
- Scientific publications and their elements, such as titles, authors, affiliations, and references, are an important data source for a quantitative analysis of science (informetrics, bibliometrics, or scientometrics).
- It requires effort to change structures in science, due to an inherent cautiousness embedded in the peer review system, the disciplinary structure, and the importance of prestige.
- Maps and other visualisations are important tools for the analysis of struc-

> tures in complex domains like science, in order to discover or locate 'hotspots', relations, communities, or trends. Maps can be made algorithmically (quantitative maps) as well as by hand (qualitative maps).

In the rest of this second chapter of Part I, we summarise our contributions in Part II and connect those publications with the research background discussed in the previous chapter. In the final section of this chapter, we list a number of overarching research themes in our work.

2.2 Publications in Part II

Our contributions in Part II can be divided in two groups. The first group, summarised in **Section 2.2.1**, consists of four publications and focuses on visualisations and static[39] structures. The first publications are on the design of an extensible[40] mapping tool and two applications of this tool. One application uses a structure presented in a map to help differentiate generic and specific natural language words and phrases, and another combines a quantitative (bibliometric) map with a qualitative mapping tool. The last paper in this group investigates non-alphanumeric characters in titles of publications, both at the general level and in major science fields. The second group of publications, summarised in **Section 2.2.2**, contains three publications and focuses on finding and analysing converging research, which are developments in which parts of science that were previously apparently unrelated, start to connect. The first two papers in this group are concerned with finding these developments without *a priori* knowledge. One paper assesses a number of different search strategies, while a second further develops one of these strategies. The last paper in this group (and final one of this thesis) does not locate converging research, but investigates the core of the knowledge base of a specific converging field (sustainability science), which exists due to a societal push for sustainable development.

2.2.1 Mapping and structures

We start in **Chapter 3** with a publication on the design and implementation of an interactive mapping, science evaluation, and bibliometric retrieval tool. In the paper we are building on work presented in Noyons (1999), but our implementation expands this work and is more general in application. As we noted in **Section 1.6.5**, a visualisation or map may be used for different purposes: domain analysis, scientometric assessment, and detecting changes in scientific structure. A user may do

---

[39] In the sense of a snapshot, not in the sense of a particular structure that appears not to change over time.

[40] By 'extensible' we mean what the Oxford English Dictionary describes as the capability "[...] of being enlarged in scope or meaning" (`http://oed.com/viewdictionaryentry/Entry/66934`, accessed 7 November 2011).

this by performing a number of tasks: discovering or locating ‘hot-spots’, relations, communities, or trends. Usually, an analysis of a research area requires all tasks and if these can be performed with a single tool, then a user can analyse more efficiently by quickly changing between tasks. So, our tool was designed in such a way in that it:

a. can be used to help with all tasks;
b. can be easily extended with additional functionality; and
c. is easily configured for different purposes.

The resulting tool has proved malleable to new applications, two of which are presented in the next two chapters. Additionally, it has also been the basis for a large number of tailor-made projects for science policy users (which we will briefly touch upon in **Section 10.1.2**).

The problem we address in **Chapter 4** (the second paper of the first group) is that some terms are more usable to refer to topics than other terms, as noted in **Section 1.3.2** on **page 26**. For example, in a field that deals with complex systems or non-linear dynamics (the example of the paper) the term ‘dynamics’ may be more useful than the generic term ‘result’. In a research area in a stable phase of development (**Section 1.5.1**), the references to common ground (**Section 1.3.4**) have also converged to a particular meaning and form. So, we adapted the interactive mapping tool presented in **Chapter 3** with functionality to visualise a distribution or spread. In our implementation, the spread is visualised by colouring elements of a map with a colour that varies between the background colour (for example light grey) to indicate low occurrence, and a dark colour (for example dark red) to indicate high occurrence (an example appears in **Chapter 10**, in **Figure 28** on **page 192**). Generic terms will occur in non-specific places on the map[41], while specific terms will only occur in specific parts of the map. This would appear either as medium-dark colouring for non-specific terms, and dark patches for specific terms. Note that the particular choice of colours can be configured.

In **Chapter 5** (the third paper of the first group) we go into the problem that words used by researchers or policy makers to describe research is not limited to the intended research, but is (as noted Sections **1.4.4** and **1.3.2**) ambiguous, fuzzy, and highly preferential. To make sense of some expression (a sentence, table, or picture), Saeed (2003, p.220) notes that people have to perform a number of tasks. In the (semiotic) jargon we introduced in **Section 1.3.1**, these tasks are:

a. infer to which objects the signs in the expression refer to[42];

---

[41] Provided that the map shows related (groups) of (elements from) publications that appear sensible and understandable in the context of the research area or field that is being analysed, instead of some random configuration of elements. See also the discussion on map validation in **Section 1.6.6**.

[42] Actually, Saeed speaks of ‘deictic’ elements. A deictic element is a language element that is contextually bound: spatial, temporal, person, or social. A simple example is the sentence “I wrote

b. fix the context in which names or other exemplars are used, possibly making a choice from a number of alternatives;
c. access the relevant background knowledge;
d. make inferences, i.e. interpret the whole expression and draw conclusions.

These tasks are also performed when assessing a map, because a user must infer what topic or research is represented by an element on the map, in the context of the research represented by the map (**Section 1.3.4**). The research we describe in this paper makes inference more explicit. We do this by looking up publications related to a user-supplied term and extract related terms from the result. Then, we enable a user to select terms from that list and to collect and structure these terms in a qualitative map. Next, we make it possible to view collected (groups of) of terms as pie-charts on a (quantitative) map of science, extending the functionality we described in **Chapter 4**. As we noted above, the occurrence of terms on an element of a map provides a specific context, which allows a user to judge whether the *actual* scientific content related to a group of terms coincides with the *expected* content. As a result, we create a way bridge the preferences in the world-view of a user with the common ground embedded in publications of a research area. Next to this application of the mapping tool, the paper also looks at how users create cognitive (qualitative) maps, by letting them create such maps of their own research areas. We find that users add more detail when expressing concepts that have a direct relation to their own research. We also find that most users prefer a map to a hierarchical structure or table. Finally, we note that users prefer to explicitly draw relations in a map, instead of only placing related elements close together.

In **Chapter 6**, the last publication of this group, we focus on the problem of 'special', non-alphanumeric characters in titles of scientific publications. We are interested in the use of these characters, the meaning associated with these characters, as well as the correlation of these characters with impact (measured by CPP/FCSm). As it turns out, the inclusion of a non-alphanumeric character has a positive effect on impact. However, this effect is not visible if the most frequent non-alphanumeric characters (the hyphen, colon, comma, left and right parenthesis) are disregarded from the calculation. Even more, in specific major fields, the effect can:

a. be positive, like in the general case;
b. be negative, i.e. including a non-alphanumeric character has an adverse effect;
c. or be absent, i.e. there appears to be no significant effect whether or not such a character is included.

We also show that the use of such characters is similar in fields with related content. For this purpose, we used another visualisation tool, a heatmap. Note that this

---

this publication last year.". In order to infer what is meant, one is required to understand what "I", "this publication" and "last year" refer to. The way to resolve this problem is to provide (or infer) a relevant context.

heatmap is not a map in the sense we defined in **Section 1.6.2**, but a visualisation of a matrix in which *both* rows and columns are clustered (Gentleman et al., 2004) and the resulting clusterings are used to reorder the rows and columns of the matrix. Additionally, by looking at specific examples, we found that some characters were used as symbolic references (see **page 116**) to concepts such as 'microinterface'. From these results we conjecture that the use of non-alphanumeric characters is stabilised in fields, up to the point that a particular character gets a specific meaning. So, if an author strays too far from this expected use, then this may influence the impact of a publication (De Mey, 1982, p.175).

For a number of reasons, this last publication of the first group forms a bridge between the first three publications and those in the second group. First, it has a focus on language: although we start by investigating characters instead of content (syntax instead of semantics, see **page 25**), we also show how such characters can start to represent content. Next to that, the publication shows that a research context, such as a field, brings about certain expectations, both with respect to content as well as format. Such expectations are also the basis for aspects we encounter in the development of converging research, which is the core focus of the second group of publications.

2.2.2 Converging science

The focus of the first group of publications was on the visualisation of structures, but we also showed that subdivisions could be useful in their own right. In this next group of papers we are mainly interested in those subdivisions of publications, and less in their visualisation, as we focus on the search and analysis of converging research (see **Section 1.5.2**, on **page 42**). We noted in the previous chapter (**page 43**) that converging research can be regarded as different common grounds that start to overlap and when successful could even start to merge, going through the different phases of disciplinarity (**Section 1.4.7**). One of the conclusions of Rinia (2007) was that a more systematic use of bibliometric methods could improve the insight in the processes, scales, and outcomes of interdisciplinary research (*ibid* p.164). Since converging research is by definition an interdisciplinary development, our search for converging research could be regarded as an attempt to improve that insight.

In the first publication (**Chapter 7**) we test three different methods to find converging research, using three possible types of context. A context is essentially a group of related papers, which can be characterised by the use certain words or phrases, names of institutions, or references (**Section 1.3.4**). Loosely based upon these characterisations, three research strategies are devised. A strategy includes at least the following parts: a way to indicate context, a way to indicate connection, and a way to differentiate between high and low growth. The first strategy is based on fields, which are essentially large groups of related papers (a macro-context, we used journal subject categories (JSCs) as proxy), and references from one field to another to indicate connections. The next strategy uses keywords (content) to characterise context, where the keywords are chosen in such a way that they are typical for a field in the first two years of analysis, and co-occurrence of keywords from different sets as connections. The last strategy used affiliation patterns (names of institutions) to characterise context, where the patterns are designed to match the disciplinary element in department names (like 'econ*' to match 'Department of economics'), and co-authors from different affiliations as connections, similar to De Bruin and Moed (1993). To indicate high growth, we use in the first strategy an absolute minimum numbers of citations. However, applying this strategy learnt us that this results in the detection of rather large developments (in number of publications). So, for the other two strategies we use another growth indicator (which we call RTG): the difference between the number of papers in the first and last year, divided by the total number of papers ($|P_n - P_1| / \sum_{i=1}^{n} P_i$). If the RTG is large, then there is a large difference between the first and last year (either positive or negative). As a result, sorting on this indicator is a useful method to find significantly growing developments (relative to the number of papers in that development). Moreover, the values of the RTG appear to exhibit a normal distribution, which makes it more convenient to assess what we can consider 'high' and 'low' (see also **page 33**). Using the citations between fields and the co-occurrence of keywords, we were able to find converging research that was also recognised by field experts. We found for example the biomaterials convergence between chemistry and material science, as well as a convergence between food science, endocrinology, and gerontology where health aspects of nutritional components are researched. Although the field of bioinformatics was deemed too old by the consulted experts, it was nevertheless detected and could be considered a good baseline result. The affiliation patterns yielded no results, probably due to the large global variation in department names, making it difficult to systematically capture institution names central to a research context.

Due to the success of the inter-field citations strategy (the first strategy we describe above) to find converging science, we continue to develop this approach in the publication that is now **Chapter 8**. The publication evaluates the previous implementation of the method and makes some adjustments. The most important adjustment is the development of another growth indicator, but we also devised a visualisation tool to compare selected growth curves (**Section 1.5.3**). Since the number of references differs largely over fields (see **Section 1.4.2**, **page 32**) we changed to an indicator expressed in numbers relative to the citing field, with the rationale that a small field which starts to cite a larger (in number of citations) field, could be a significant development from the point of view of that citing field, while it might not be significant from the point of view of the larger, cited field. The resulting indicator we call the 'growth rate g'. This indicator exhibits a log-normal distribution which we use (analogous to how we used the distribution of the RTG) as a more 'objective' basis to differentiate between high and low growth. Using our adjusted search strategy, we again found the bioinformatics convergence, the baseline result we mentioned above. Added to that, we found (among other examples) a convergence between the fields *genetics* and *communication* on ethical consequences of research in genetics; research on biodiesel that appeared to involve mostly Indian institutions; and the 'econophysics' convergence between *physics* and *econometrics*. Also, we found that some developments were only in one direction, and could thus not be called convergence. Additionally, we noted that some developments showed concrete causes, such as the introduction of a new instrument, method, or societal interest, while other developments had a more generic cause, such as an increased interest in statistical methods in bioinformatics. Finally, we assessed different elements related with publications to indicate *coherence* (see **page 42**) and found that for this purpose the spread of publications over (different) journals, which should decrease as a research context is stabilising, appears promising.

Our final paper in **Chapter 9** continues with the theme of converging research, but instead of trying to use structures to locate change, we now analyse structures in a field which is known (or said) to converge. More particularly, we turn to the problem of analysing the disciplinary roots of the interdisciplinary (converging) field 'sustainability science'. To delineate (see **Section 1.4.4**) this field, we extracted all publications with titles that match the pattern "sustainab*". This works because the field is still relatively young and (still) needs a recognisable marker[43]. From this delineation we extracted a core knowledge base (KB) of highly-cited publications, *exemplars* (**Section 1.3.4**) of the converging fields. We were able to show that this core KB of sustainability science can be divided in three parts, which correspond to research into the social ("people"), economic ("profit"), and environmental ("planet") aspects of the problem. Finally, although the field is *conceptually* converging[44], from an analysis of the citations between the parts of sustainability science, we can not yet see signs of an increasing convergence between the different parts of the field. So, although it might be desirable that in sustainability science the interdisciplinary approach becomes standard, this appears not yet achieved.

2.3 General research themes

From the above summary of our publications in Part II, we can distill some of our overarching research themes and interests.

- The first theme is the focus on structures, an abstraction which we have specified as a collection of sets of (elements from) publications. We visualise structures in a map and use characteristics of (potential) structures to find

---

[43] An important reason this marker is required is that, until recently, only a handful of specialised journals existed for this field. As a result, publications which did not fit the specific nature of these journals, had to be published in other journals, which had no *direct* relation to sustainable development. In a specialised sustainability journal, a marker would have been unnecessary, because the context of the journal is implicitly stating this application and purpose, as opposed to non-sustainability journals.

[44] By which we mean that it is widely accepted that addressing problems in this field requires a concerted, interdisciplinary approach.

converging research.
- Another theme is the focus on the cognitive aspects of science, something we already noted in the introduction. This is most notable in Chapter **5**, but the focus on language as a tool for communication of knowledge can be considered part of this theme as well, which is present in most contributions.
- Closely related with the cognitive focus is the theme of research context, or common ground. It is an underlying assumption of the use of maps in Chapters **4** and **5**, but also in the search for converging science in Chapters **7** and **8**.
- The use of jargon, exemplars, and other elements such as institution names is another theme related to research context and common ground, which we most visibly applied in the search for converging research in Chapters **7** and **8**.
- Visualisation is of course another important theme. In particular the mapping of research in Chapters **4** and **5**, but also in Chapter **9**. Added to that, we also used charts and plots in many publications, as well as a heatmap in Chapter **9**.
- Also a theme is the fundamental gap between tangible and intangible aspects of science. This is most apparent in Chapter **5**, in which we describe a tool to create a 'translation' between tangible and intangible aspects.
- Building tools and methods is also a theme. Clearly, this is the focus of Chapters **3** and **5**, but also the research on converging research in Chapters **7** and **8** mostly deal with the development and fine-tuning of search strategies.
- Finally, the generic character of these methods and tools, as well as their application on large sets of publications, as opposed to the application on a limited, focused set of publications, can also be considered a theme in our research. This is most apparent in the search for converging research in Chapters **7** and **8**, as well as the investigation of non-alphanumeric characters in Chapter **6**.

These themes can be summerised as the search to increase the applicability and usability of quantitative analyses of science. By applicability we mean that a particular analysis or tool helps to address the problem at hand, similar to the utility aspect of map validation introduced in **Section 1.6.6**. By usability we instead refer to the ability of users to use or handle an analysis or tool and apply it to a problem; so, usability can be regarded as a precondition for applicability. For example, even though a map may be a good tool to analyse some research field, users may still not be able to apply it due to lack of knowledge of the tool, or a disconnection between world view and map. Improving usablity (and thus applicability) can be regarded as the overarching theme of the first three chapters of Part II, where we improve both the mapping tool and try to connect to the world view of users. However, applicability may be more general as well and refer to the applicability of quantitative analyses in general. Our work on non-alphanumeric characters in titles of scientific publications (**Chapter 6**) is an example, in which the applicability is in that we show

how the format of titles are also part of the common ground of a field, and should therefore be taken into account when participating in the research of that field. The chapters on converging science also contribute to applicability and usability. The research on methods to find converging science in the entire WOS in Chapters **7** and **8** are not optimised for a particular field, but applicable to the whole WOS database. Added to that, in **Chapter 9** show that quantitative analyses may also be applied to more abstract developments like sustainability, where there is only an indirect link to research.

We divide users into three main groups: 1) Researchers active in a field being analysed; 2) researchers, policymakers, or other interested people familiar to a field, but not subject to analysis; 3) a general audience. Work we have done on visualisation and mapping has mostly been targeted at the first two groups. The work on searching for converging science is mostly of interest to the second group, while the work on sustainability science is specifically targeted at the researchers active in this emerging field. Finally, the work on non-alphanumeric characters is targeted at a general audience. Of course, the people in these main groups are not homogeneous. For example, policymakers exist at different levels, ranging from ministers or directors of national science foundations to their advising personel. At these different levels, the questions change, and subsequently the requirement for detail and scope change also: from general questions requiring broad scope and less detail, to more focused questions involving more restricted scope and more detail. Most of the work in this thesis is malleable with respect scope and can be given an appropriate (or required) level of detail.

References

De Bruin, R. E. and Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics, 26*(1), 65–80.

De Mey, M. (1982). *The Cognitive Paradigm.* London: D. Reidel Publishing Company.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology, 5*(10), R80.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool.* PhD thesis, Leiden: Leiden University.

Rinia, E. J. (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer.* PhD thesis, Leiden: Leiden University.

Saeed, J. I. (2003). *Semantics.* Malden (MA): Blackwell.

Part II

Published Papers

3

Improving the functionality of interactive bibliometric science maps[*]

Abstract

The use of a map as a metaphor of a scientific field is an established idea and using it as an interface to bibliometric data seems to have great potential. Nevertheless, our own implementation of such an interface came up with some limits inhibiting the user to comprehend as to what he was looking at. As a result, the map was not used to its fullest potential. The implementation described in this paper as a high-level (conceptual) design, addresses the problems noted by users. It combines both top-down and bottom-up access to the bibliometric data, something we see as vital to mapping internal knowledge onto the external depiction and vice versa. And as such, it becomes a more complete tool to explore the mapped scientific field and to find and retrieve relevant information.

3.1 Introduction

Maps of science are a metaphor to abstract from the intricate details of the innards of a scientific discipline. The metaphor can nowadays be regarded as an established one since work on it started at least three decades ago (and maybe even before that) and has steadily been used and developed since (cf. the comprehensive review on visualisations).

When interested in a scientific discipline,[45] or more specifically in the relations existing between ‘research participants’ (White and McCain, 1997) or actors (authors, organisations, countries, etc.) within such a discipline, a picture presenting a spatial

---

[45] The definition of which may be through keywords, journal titles or even benchmark actors.

---

[*] Originally appeared in Scientometrics 51(1), 55–68, 2001.
Reproduced with permission of the publisher.

impression of (groups of) associated elements has shown to be at least a promising presentational aid. In evaluative bibliometric research they are used as a domain analysis tool, usually hinting on their policy-supportive capabilities (Healey et al., 1986; Hinze, 1997; Noyons, 1999; Noyons and Van Raan, 1998 and Schwechheimer and Winterhager, 1999). Where standard discussion in a domain analysis creates lists of indicators, which are essentially answers to questions like 'What are the reigning topics in the field?' or 'Who are the main actors in a certain area?', the maps then add the ability to indicate how the main topics or actors relate to one another by showing where these elements are located.

The increasing availability of computers has enabled researchers to move beyond the paper presentation by stirring different flavours of interactivity into the maps. These interactive maps have been especially prevalent within the field on Information Retrieval (IR) (Brooks and Campbell, 1999; Kohonen, 1998 and Lin, 1997). The roots of these efforts already date back to the 1980's or maybe even the mid-1970's (cf. (White and McCain, 1997, p.104)).

The efforts in bibliometrics however are perhaps a bit less diverse, especially those available on the WWW (Chen, 1999; Järvelin et al., 2000 and Small, 1999). In (Noyons and van Raan, 1998) the outline is set for our addition to interactive maps intended for science-supportive purposes. Those maps are essentially pictures in which hypertextual elements (HTML) are embedded in order to become interactive. One of our main challenges has been to further develop the functionality supplementing these maps in order to make these useful tools in addressing policy-related questions using bibliometric data.

The paper is organised as follows. First, we further explore our motives to build a new tool. After that, we touch upon the high-level design and present a (preliminary) version of the implementation. Finally, we present a short example of a science-policy question and a way to aid resolving that question using the interactive map and its functionality.

3.2 Objective

The use of hierarchical clustering algorithms to group elements such as words and citations based on co-occurrences is somewhat of a tradition within the field of bibliometrics – though not one without criticism. In this paper we choose to ignore such a discussion altogether and simply assume a picture built with these ingredients provides a useful and usable structuring of a scientific discipline. We can do this since in this paper we want to present the functionality supplementing the map, which depends in no way whatsoever on the brick and mortar used to create the map.

The target of this paper is to present our next-generation interactive map and to hint at questions and ways to resolve these using the interactive map. This makes our target audience both the technically interested researcher as well as the curious policy-user interested in new ways of applying bibliometrics.

3.3 Motivation

3.3.1 Maps as interfaces

Our maps are based on co-occurrences of phrases, a technique which is essentially an extension of the mapping based on co-occurrences of words, which (in a bibliometric context at least) can be traced back to (Callon et al., 1986). Nevertheless, a science policy maker is usually not primarily interested in frequently occurring phrases, which are the elements our co-word maps are made of. Instead, the map is intended to provide some kind of interface to the actual bibliometric data.

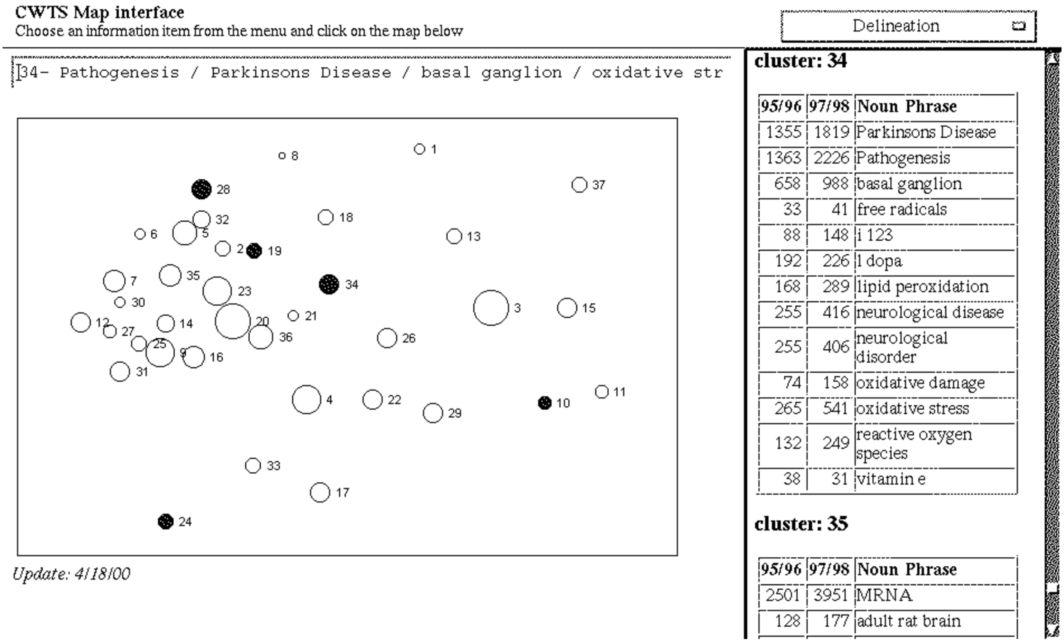
Figure 3 shows an example of one of our HTML-based maps. It shows on the left the virtual landscape, which is a map of the 'neuro-sciences'. The clusters are represented as dots with a certain size and colour, both of which represent additional information (number of publications and growth of the number of publications, respectively). The distance between the clusters represents the degree of association with respect the number of co-occurring phrases.

The representation in these maps is deliberately simple, and not to be compared with sophisticated 3D visualisations like the ones in e.g. (Hemmje, 1995). It is essentially just a 'point- and-click' interface: the clusters can be clicked on using the mouse, resulting in the display of the information relevant to that specific location on the map and chosen topic of interest (focus). In **Figure 3** for example, the user has just clicked on the central dot bearing the number 34", and the title of this cluster is shown in the text field above the map. The list on the right called 'Delineation' is the current focus, as requested by the user. It is a list of the delineating noun-phrases, i.e. the noun-phrases that were grouped together by the clustering algorithm. For every cluster in the map, a link is created to the sublist relevant to that specific cluster and embedded in the picture. After the click on the cluster '34' for example, the list on the right side of the map changed as to display the sublist with the title 'cluster 34'.

The box with the word 'Delineation' is actually a graphical element called a 'choice box'. And by manipulating this box, the user can choose another focus, like the 'Most cited authors' or the 'Most active countries'. The newly chosen theme corresponds to a new list and the links embedded in the map will be adapted as to contain links to the sublists in this new list of topical information.

3.3.2 Critique

Interactive maps such as the one shown in **Figure 3** have been developed and used and by us for some time now. The use of the maps has been successful in the sense that users indicated they could both relate to the depiction as well as saw the potential usefulness of the interface and its functionality. However, through discussions with these same users certain limitations were noticed as well. Those limitations did not concern the use or validity of maps as such, but more specifically the functionality



CWTS Map interface
Choose an information item from the menu and click on the map below

34- Pathogenesis / Parkinsons Disease / basal ganglion / oxidative str

Update: 4/18/00

cluster: 34

95/96	97/98	Noun Phrase
1355	1819	Parkinsons Disease
1363	2226	Pathogenesis
658	988	basal ganglion
33	41	free radicals
88	148	i 123
192	226	l dopa
168	289	lipid peroxidation
255	416	neurological disease
255	406	neurological disorder
74	158	oxidative damage
265	541	oxidative stress
132	249	reactive oxygen species
38	31	vitamin e

cluster: 35

95/96	97/98	Noun Phrase
2501	3951	MRNA
128	177	adult rat brain

Figure 3 An example of a clickable map with associated list.

provided together with these maps.

The functionality supplementing these first-generation maps is limited since the maps are, in essence, just hypertext: pre-set navigational aids (the links) through a body of information. And although we may supply many of such pre-set routes, each with a slightly different focus, there are limits as to what the user can do with them. For example, selecting two clusters on the map and combining their data, or the concurrent use of different lists (e.g. ‘Most cited organisations’ together with ‘Most active organisations’) are hardly possible to implement without going extensively beyond the basic (static) framework.

3.3.3 Mapping the map

The described inflexibility in accessing the data is also an important factor in the trouble the users expressed in comprehending the map. Again, not because they question the validity of the presented structure or the validity of using such a map as a science policy tool, but because they find it difficult to map their own ‘internal’ structure, ‘domain theory’ (Healey et al., 1986) or ‘mental maps’ (Tijssen, 1992) onto this ‘external’ structure.

Here, the common word ‘structure’ can be somewhat misleading since it suggests the two may be related. This is only very indirectly so since both structures are of a completely different kind: the one the user is looking at is a two-dimensional or perhaps three-dimensional picture, while the internal one will not be a picture or map at all. But besides this obvious difference, both user and picture have attributed degrees of association to pairs of elements. And the existence of these associations is what we as scientometrician (mapmaker) would like to exploit in order to bridge the gap between the presented structure and the knowledge the user already possesses.

As an analogy, assume someone is presented an aerial photo of his neighbourhood. At first, he will most likely not recognise what is shown on the picture. To change this, one of the first things he will probably do is to search for familiar landmarks, such as churches, parks or maybe certain crossings. When found, these landmarks and their relative position will enable the user to gain knowledge as to what the other elements on the photo represent, and finally (as a result) the photo as a whole. Using this analogy, we argued that by providing the user with the functionality to enable him to search for familiar elements on the map, he could become familiar with that map.

Our map has (just like any descent map) labelled the elements it displays. Nevertheless, our previous studies showed this labelling is not enough to bridge the 'mental gap' we mentioned above. So, when presented a map like the one in **Figure 3**, the user should be enabled to compile a list of e.g. his publications, or his organisation together with some benchmark institutes. Having compiled such a list, the elements should somehow be positioned on the map. Displaying coloured numbers representing the total number of papers produced during a given period with every colour representing another actor, may for example already accomplish this. Comparing this display with his internal structure may now reduce the metal gap, thus increasing the user's comprehension as to what he is looking at.

Note how the search for elements (i.e. compiling lists of familiar elements) in a map goes from the data to the map. This 'facts-to-map' way of accessing and displaying information is what we dubbed *bottom-up* access, as opposed to the *top-down* or 'map- to-facts' access provided in our first-generation maps. To summarise, it is the inflexibility of the framework we use to add interactivity to our maps that hinders the maps to be used to their fullest. Extending the framework by adding bottom-up access might well be possible, but this would be a patchwork solution. Thus, we looked beyond hypertext.

3.4 Design and implementation

The new implementation can be regarded as a direct extension of the HTML-based one, although it is intended to replace the latter. This is one reason why we elaborated quite extensively on the old one. Another reason is that we can now indicate (in the discussion below) on which points we think we made progress in functionality with respect to the previous implementation. Again, this new implementation is not innovative in the sense that it has never been done before – and we do not claim it to be. However, we do not know of a readily available alternative with the specific target use (policy-support) that we could actively deploy on the WWW and also conformed to the following (design) criteria:

1. the interface should be able to provide all functionality currently available in the HTML-based clickable maps;
2. it should provide means to access the information bottom-up and display this on the map;
3. it should be able to combine bottom-up with top-down retrieved information;
4. it should be able to retrieve and aggregate data without any structural limits.

With the latter is meant that the application should only limit combination and aggregation for practical reasons, e.g. to avoid excessive use of the computer's processing time or its memory. Actually, the second and third criterion could be regarded as a special case of the fourth criterion, with such limits implicitly in place.

3.4.1 Conceptual design

The maps are used to access and display (bibliometric) data. For the new implementation, we assume the data are stored in more general sources like databases, instead of just lists of hypertext. But what is actually available in the data? In other words: who and what is taken from the Universe of Discourse, sampled and perhaps already aggregated to become stored in these databases? To answer this, we made a formal distinction between the publications and the publication-producing[46] entities. This distinction is a matter of semantics: the information about the actors (a paper-producing entity) is in our line of research usually accessed through the associated (produced) publications.

In **Figure 4** is tried to sketch the conceptual design, showing the major components and some relations between elements.

The design is divided into two parts. To the left is shown the graphical user interface (GUI) with the *desktop* as central metaphor. This part represents the interactive part of the application. On the other side is shown the informational body, with the entities and the maps, both of them contained in data sources. Since this informational body changes from application to application, we have used the *project* as the binding metaphor.

3.4.2 Project

The project embeds the application-specific details regarding data sources, form and

---

[46] This 'production' of papers is to be taken broadly and is certainly not to be interpreted (per se) as 'writing'. Instead, it should be thought of as 'there exists some kind of association between the entity, and one or more publications'. The entities may have indeed written it, as authors do; or they may have employed the authors who wrote them, as organisations do. And when we look at it this way, even the publications themselves produce papers, with the publications they produce given a special name: references.

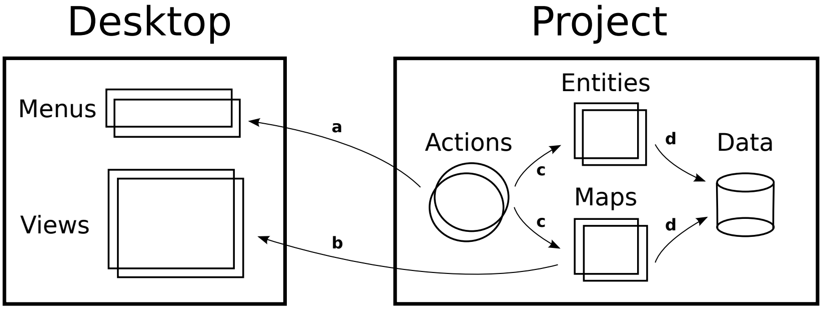


Figure 4 The conceptual design of the application with some relations: a. listed-in; b. displayed-in; c. parametrized-by; and d. contained-in.

content of publications and publication-producing entities. Some of these details are encoded in the set of actions that are available to the user to execute. The complete set of actions makes up the whole of the functionality of the interactive map. More on actions is said below.

3.4.3 Desktop

The interactive, graphical part of the application is modelled after the familiar desktop, which has been the prevailing graphical user interface metaphor for well over twenty year now – predating even the current PC-era which has promoted the metaphor so fervently. Therefore, the science policy maker or researcher using the application should already be familiar with the graphical elements used in this desktop.

The desktop contains one or more elements called views. These views are the more general replacements of the clickable maps. When the user clicks on an element in the view, that element will become selected. If the element were a cluster displayed as a dot (just like in **Figure 3**) for example, the dot would change colour or change shape as to indicate it is currently selected. Would the user click on the cluster again, it would become unselected and its colour or shape would be restored. Clicking in a view, but not on a cluster, enables the user to select a group of clusters by dragging a rectangle around them. All this allows the user to indicate specific areas of interest instead of single clusters.

The desktop also contains menus, which are a more general replacement of the choice box shown in the upper-right corner of **Figure 3**. These menus list the actions available to the user and are contained (as would be expected) in a menu bar at the top of the desktop. Where available and if appropriate, the limited scope or context of an action can be highlighted by showing (popup-) menus on a cluster or view. This should also be a familiar thing to many users of modern computer systems.

3.4.4 Actions

In the HTML-based implementation the user could:

- change the focus of the information listed next to the map;
- load a different picture of the map with new elements on it, such as links showing e.g. the more significant citation traffic between the cluster clicked on and other clusters;
- load a dynamic version of the map, showing an animation of the change in position of the clusters over a number of years;
- load a different map altogether, with e.g. a more detailed map of elements from a lower part of the clustering hierarchy.

The new implementation tries to generalise these actions into four classes that will be described below. The first three of these actions are the information-related ones. For these we can distinguish between two 'targets':

- the intra-field target: information aggregated to a specific cluster, a group of clusters or complete map, e.g. to get at the impact of a specific actor within that map or group of clusters;
- the inter-field target: information spread over two or more clusters or map, e.g. the impact of a specific actor in each cluster separately, or the citation traffic between a specific cluster and the other (selected) clusters.

Note that there is thus a difference in retrieving a list for a whole map and retrieving a list for all clusters in a map. The first is an example of an intra-field target (the field is the whole map) and the second of an inter-field target. For example, the top five most active organisations would in the first case result in a list of five organisations, while in the second case it would result in a list of n times five organisations, where n is the number of clusters in the map.

The actions in the HTML-based implementation have been generalised into the following four classes.

A. Retrieving precompiled statistics.

This first class of actions enables the user to answer questions about the field as a whole, such as:

- What are the fields having the most impact?
- What are the citation characteristics of the field represented by this map?

Therefore, the access provided by these actions information is purely top-down and their target is always intra-field with the field equal to the whole map.

B. Retrieving precompiled statistics limited to specific (lists of) entities.

This class of actions was actually (and implicitly) the default in the HTML-based implementation: the user selected a cluster, which 'limited' in a sense the list shown on the right by changing to the sublist specific to the cluster just selected. In the new implementation the user can limit the lists of information by:

- Selecting groups of clusters.
- Selected parts of other lists, e.g. the selected top five most cited organisations.
- A search result, e.g. the organisation the user is working for together with a couple of familiar ones.

Questions asked could thus be like 'Show the activity of the most cited organisations in clusters A, B and C as well as those in my search result.' Usually, actions in this class have the first type (just listing precompiled statistics) as a special case.[47]

C. Creating customised lists.

This class of actions is the result of our design criterion four, which dictates there should be no structural limits as to what to combine and aggregate. However, this does not imply every project has implemented this ability, mainly due to the already-mentioned resource constraints.

D. Miscellaneous actions.

This class of actions is not data related, like for example the opening and closing of views, rotating and scaling of maps or loading of new projects.

The results of the information-related classes A, B and C are lists of data representing instances of one or more entities, usually together with counts or other statistics. These lists can either be shown as a list, or where applicable as bar charts or pie charts (see also **Figure 5**).

As said above, the set of actions supplied in a project should enable the user to answer his questions regarding the field. This implies the mapmaker should compile the actions with care.

---

[47] Nevertheless, not always. Take for example the field citation mean (F_{CSm}), an indicator employed at the CWTS that applies to specific fields only partially related to the field shown on the map. There could be more than one F_{CSm} for a specific map, corresponding to different field definitions that overlap the one defined through frequent phrases that is shown on the map. Nothing in the map could limit this list.

3.4.5 Implementation

When we write this, a Java implementation of the application is in its final stages. This specific language was chosen because it enabled us to let the science-policy maker or researcher in the field to continue accessing our work through the WWW. We will provide a link to a demonstration on our homepage. **Figure 5** shows a preliminary version of the java implementation[48]. In this figure, the desktop contains one view of a map of the field 'Complex systems'. The user is looking to the view at 99% of its default size, but rotated it 203 degrees. This particular map has been created for the year 1997 and contains 10 clusters. These clusters have been labelled with both a number and a title. Note how one cluster (number '9') has been dragged from underneath the large central cluster (number '8'), which is indicated by a small line connecting the cluster to its original position below the central one.

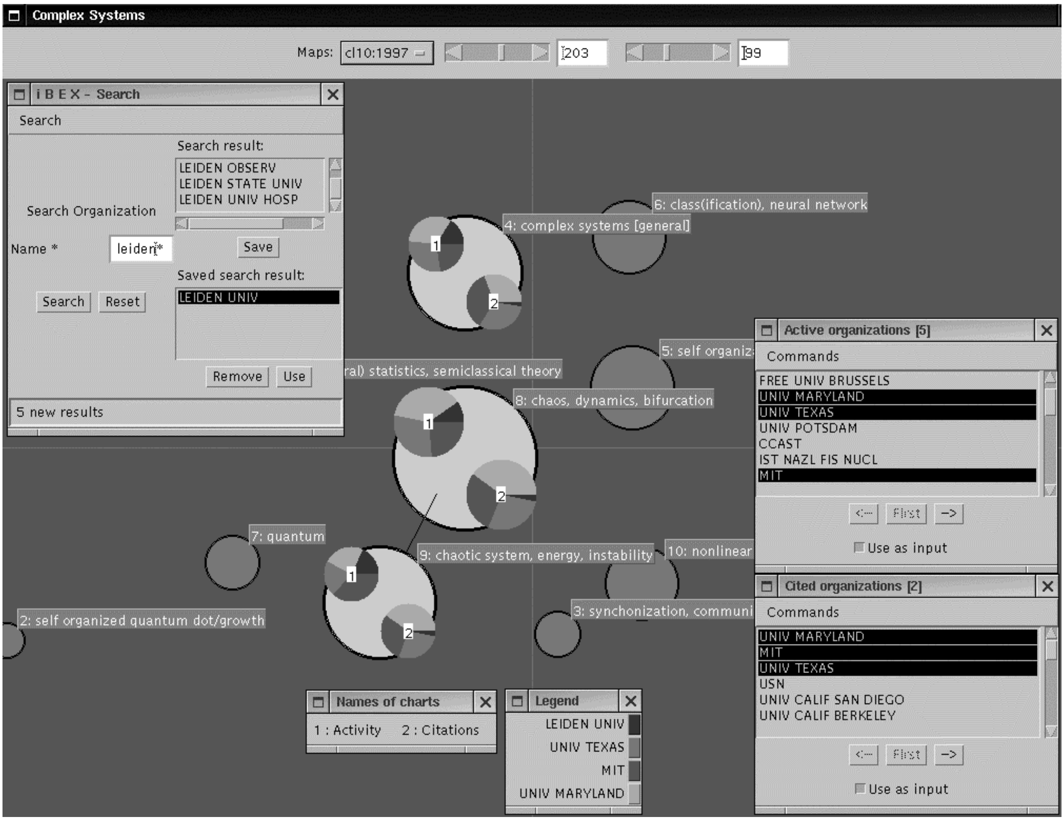
The three central clusters have been selected, which is indicated by giving them a different colour. For these three selected clusters, both the most cited and the most active organisations have been retrieved (i.e., an *intra-field* target). The organisations showing in reverse video have been selected for display as pie charts, one chart on each of the selected clusters (now the target is *inter-field*). The user has also searched the database for an organisation, using the search term `leiden*`. This search resulted in a list of candidates from which the one indicated as 'LEIDEN UNIV' has been saved as the one to keep in the search result. In this case, the search result is a single instance, but the implementation can also handle a list of instances as the search result. For this search result, both citation and activity data were retrieved (not shown in the figure). This bottom-up accessed part of the bibliometric data has subsequently been combined with the data accessed top-down and is now also displayed in the pie charts. To aid the user, the two different pie charts have been given a number and their descriptions are shown to the left of the legend.

An example of answering a science policy question

As an example of use, we take the following question and show how it could be answered using the application. Since the question is rather general there can be given more than one answer. The question is the following: 'Who uses the European knowledge base?' In order to begin to look for an answer to this question, it must be a bit more specific. We could restrict the 'who' somewhat by reformulating it as: 'What are the most active organisations using the European knowledge base?' The second part of this question ('the European knowledge base') is still very general, and is a good example of how the map may refine and target this vague notion. For example, we could select all fields on the map and retrieve the number of citations for the country (aggregates) EU, Japan and US and have the result displayed as pie charts on the map. Based on this we could select the areas showing a (relatively)

---

[48] We have baptised this java implementation 'iBEX', which is both an animal as well as an acronym that stands for 'Interactive Bibliometric EXplorer'



Maps: cl10:1997 203 99

Name * leiden*

Search result:
LEIDEN OBSERV
LEIDEN STATE UNIV
LEIDEN UNIV HOSP

Saved search result:
LEIDEN UNIV

5 new results

Active organizations [5]:
FREE UNIV BRUSSELS
UNIV MARYLAND
UNIV TEXAS
UNIV POTSDAM
CCAST
IST NAZL FIS NUCL
MIT

Cited organizations [2]:
UNIV MARYLAND
MIT
UNIV TEXAS
USN
UNIV CALIF SAN DIEGO
UNIV CALIF BERKELEY

1 : Activity 2 : Citations

LEIDEN UNIV
UNIV TEXAS
MIT
UNIV MARYLAND

Figure 5 A screen shot of an early version of the application.

large number of citations for the EU. Now we have implicitly refined the question to be: ‘What are the most active organisations in the areas where the EU is relatively strongly cited?’ To look for an answer to this last formulation of the question, the user could retrieve a list of the most active organisations in the now limited selection of clusters. The result of this example can be regarded as a partial answer to the original question (with the world limited to the world of the displayed map, of course).

3.5 Conclusions

The objective we set out with when starting the design of the presented implementation was, simply put, to overcome problems noted with the use of our first- generation interactive maps. Especially the inability to locate a familiar item on the map was a reason to re-implement our interface. In this, we have succeeded with the presented design. Projects are already planned that will actively deploy the new implementation to enable users to analyse a specific domain, and we can start work on refining the implementation and design.

Acknowledgements

An important part of this work was done as part as research sponsored by the EC (TSER programme). We thank the S&T 2000 organisers for providing the opportunity to publish this article, as well as the referees for their comments on an earlier draft of this work.

References

Brooks, M. and Campbell, J. (1999). Interactive graphical queries for bibliographic search. *Journal of the American Society for Information Science, 50*(9), 814–825.

Callon, M., Law, J. and Rip, A., editors (1986). *Mapping the Dynamics of Science and Technology.* The MacMillan Press Ltd, London.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management, 35*(3), 401-420.

Healey, P., Rothman, H. and Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233–251.

Hemmje, M. (1995). LyberWorld: a 3D graphical user interface for fulltext retrieval. In *Proceedings of the ACM/SIGCHI'95*, pages 417—418.

Hinze, S. (1997). *Mapping of Structures in Science and Technology: Bibliometric Analyses for Policy Purposes.* PhD thesis, Leiden: Leiden University.

Järvelin, K., Ingwersen, P. and Niemi, T. (2000). A user-oriented interface for generalised informetric analysis based on applying advanced data modelling techniques. *Journal of Documentation, 56*(3), 250–278.

Kohonen, T. (1998). Self-Organization of very large document collections: State of the art. In *Proceedings of ICANN98*, pages 65—74.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40–54.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool.* PhD thesis, Leiden: Leiden University.

Noyons, E. C. M. and Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics, 41*(1-2), 61–67.

Noyons, E. C. M. and van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

Schwechheimer, H. and Winterhager, M. (1999). Highly dynamic specialities in climate research. *Scientometrics, 44*(3), 547–560.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.

Tijssen, R. J. W. (1992). *Cartography of Science: Scientometric Mapping with Multidimensional Scaling Techniques.* PhD thesis, Leiden: Leiden University.

White, H. D. and McCain, K. W. (1997). Visualization of literatures. In Williams, M. E., editor, *Annual review of information science and technology (ARIST)*, pages 99–168. Information Today, Inc. for the American Society for Infomation Science.

4

Using Bibliometric Maps to Visualise Term Distribution in Scientific Papers*

Abstract

Bibliometric maps are scaled representations of clusters of papers and provide an overview of a scientific field as provided by those papers and the references between them. We have enhanced our interactive bibliometric explorer (iBEX) to visualise the distribution of phrases over such bibliometric maps. Assuming that generic (non-content-bearing) phrases occur in most of the papers in the map, those phrases confined to a specific part of the map will probably be more content-bearing than those distributed all over the map. Such knowledge can eventually be used to evaluate a selection of words and phrases.

4.1 Introduction

The use for bibliometric maps (BMs) is the same as that for an ordinary street map: both are intended to guide the reader through a landscape. For an ordinary street map this landscape will be something very real. But for a BM the depicted landscape is a totally abstract and for some readers this landscape will be an unfamiliar one. Therefore, we first provide a short introduction into bibliometrics and bibliometric maps. After that, we will sketch our motivations behind this endeavour. Then we will describe a sample analysis with some early experiences with the application. And finally we will present some conclusions.

4.1.1 Bibliometrics

A bibliometric analysis starts by collecting a set of papers published in journals,

---

* Originally appeared in the Proceedings of the Sixth International Conference on Information Visualisation, 697–202, 2002, ©2002 IEEE, All rights reserved.
Reproduced with permission of the publisher.

proceedings or other types of peer-reviewed paper collections. This is usually done with the use of expert knowledge. For example, someone working for an organisation involved in a bibliometric analysis is asked to provide a set of keywords or journals which describe the scientific field his organisation is active in.

Depending on the granularity of the analysis, papers are associated with entities like authors, organisations or countries. These entities are commonly referred to as "actors". Relevant statistics in such an analysis would (among others) be the activity (number of papers) of an actor, as compared to the total output of the scientific field, as well as the number of citations an actor has received, compared to the number of citations that a paper is likely to receive in the scientific field at hand (Moed et al., 1995). The analysis may additionally include a list of benchmark actors or other additional statistics to compare the actors subject to analysis more clearly.

4.1.2 Bibliometric maps

A bibliometric map is a scaled, two-dimensional representation of a clustering of attributes of scientific papers. More specifically: using a selection of attributes, we compile an association matrix. This matrix is used as the basis for an hierarchical clustering algorithm. The hierarchy is cut off at a specific point according to some well-defined criteria. The resulting clusters are then used to create a distance matrix, which is mapped onto a two-dimensional space using a Multiple Dimensional Scaling algorithm. The result is a list of clusters and coordinates in a two-dimensional space that makes up our map.

It is important to note that only the distances between the clusters in the maps bear any meaning: those clusters mapped closer together are likely to have more attributes in common than clusters placed further away from each other. More specifically, there is no meaning attached to the axes, as might be the case in other kind of displayed material.

There are different types of maps, corresponding to different attributes of papers and different ways of creating association matrices between these attributes. For example, we could make a map using the most-cited actors, where the actors are associated by the number of times they appear together in the reference lists of papers (see for example Small (1999)). Alternatively, we can make a map using a list of frequently appearing words or phrases taken from the title or abstract of the papers (as pioneered in Callon et al. (1986)). In our work, we have concentrated on the latter type of map, using Noun Phrases (NPs) (Noyons, 1999) instead of just (single) words. Such NPs are phrases taken directly from the title or abstract of the paper, but extracted (and normalised) using a parser for the English language.

Although the maps may have been built using association matrices of selected paper attributes, they (naturally) also provide an overview of the papers from which the attributes were extracted. And therein lies their use: they provide a condensation of the set of papers and can be used to e.g. provide nuance and additional information by displaying related structures closer together. And as such, they can be a great additional tool to enhance the comprehension (and as a result, their use) of bibliometric analyses.

4.1.3 Common usage and interactivity

The maps we create are often used to support science-policy decision making. Users required to make such decisions need a tool that enables them to retrieve details about the structure of a scientific field, in order to properly support their policy (Van Raan, 1999). They usually have a good knowledge of the coarser structures in the field and bibliometric maps enable them to use this knowledge to get at these finer details.

Partly due to this demand, we have started to create and deploy interactive maps (Buter and Noyons, 2001). These interactive maps have become a complete application which we have baptised iBEX.[49] With this application we have created an Information Retrieval tool that is able to retrieve different combinations of bibliometric statistics and other data based on selected clusters in the map, or the map as a whole.

The application also enables the user to search for specific elements, like journals, authors, institutes or words, and show the associated papers (or their citations) as numbers or charts on the map. This allows the user to become acquainted with the abstract landscape depicted before him by looking for familiar elements – just like one would do on that ordinary street map.

4.2 Motivation

4.2.1 Noun phrases and bibliometric maps

As already mentioned, there are a several types of bibliometric maps besides the one we create using NPs. But there are a number of advantages in using terms taken directly[50] from papers, such as:

- they provide the most up-to-date and direct connection to the scientific activity described in the paper;
- they are biased only to the jargon or language usage they are a part of, and are less likely to be be influenced by social (co-citation) structures;
- since they consist of normal words, they should (up to a certain point) be

---

[49] This is both an acronym of Interactive Bibliometric EXplorer as well as a robust breed of wild goat from the Alps and Apennines, with large ridged recurved horns.

[50] Here, by directly we mean as opposed to e.g. database keywords which may not directly appear in the text.

accessible to non-experts.

4.2.2 Noun phrases selection

On the other hand, a disadvantage of the use of NPs is the sheer abundance and variability of different words used in papers. So, before NPs can be used to cluster papers, they have to be submitted to some sort of selection process. The objective in this selection process is one common to a lot of such processes: to select a set of elements (NPs) that has a nice balance between precision and recall. After all, a low precision (but high recall) may result in a cluster structure that combines too many (relatively) unrelated papers; whereas a too high a precision (with low recall) may result in many small clusters. And added to that, since we are trying to visualise a scientific field, we want to be as complete in our description as possible; and a low recall would limit that completeness and (as a result) the usability.

Developing such a selection process, even for a well-defined, specific task, is an open research question (and one we are actively pursuing). A commonly used criterion to get a grasp on things is that generic (non-content-bearing) words are distributed randomly (evenly) over a corpus, whereas content-bearing words are only used in a limited set of (semantically) related papers (Kim and Wilbur, 2001). Since a BM consists of clusters of related papers and displays related clus- ters closer together, one would suppose that content-bearing terms (words, phrases) are mostly contained in clusters in a specific region of the map.

The use for such a visualisation is twofold. First, it provides a way to gain insight into the spread of a specific term or list of terms. And second (as a result) this can be used to evaluate a list of terms (that may have been generated automatically) on its appropriateness as the basis for a BM.

4.3 Sample analysis

We enhanced the ability of our bibliometric explorer to display graphical material like charts or numbers on clusters in the maps, with some basic functionality to visualise the distribution of a list of phrases. Such a spread simply associates each (selected) cluster with an integer $n \geq 0$. This integer represents the number of (distinct) papers containing one of the selected NPs. This number is plotted on each cluster and the colour of the cluster is changed: clusters containing close to zero papers will fade into the background whereas the clusters containing a relatively high number of papers are given an increasingly contrasting colour instead.

4.3.1 Map creation

The BMs we used to visualise the term spread are not built using words or NPs. After all, this would require a selection process again, with all the subjective bias we would like to avoid. Instead, we have created maps based on what we may call journal citation'. That is, the measure of association used to build the first matrix in our map-creation process, was the amount to which the journals of the selected papers refer to (papers in) journals in the reference lists of these papers. For example, a paper appearing in journal A makes reference to papers in journals B, C and A (a self citation, so to speak). As a result, the journal the associations (A,B), (A,C) and (A,A) are added to our matrix.

The papers we selected for this project, all have to do with "complex systems" and "non-linear dynamics".[51] We created maps for each of the publication years 1997 up to 2000. So, in a specific map, the clusters contain only (journals of) papers published in a single year. Every publication year has about 3000 to 3500 selected papers. The example visualisation in the figures below, all use the map with the papers published in 2000.

4.3.2 Map usage

Figure 6 provides an overview of our application. The clusters in the maps are represented by filled circles, the area of which corresponds to the number of papers in that cluster.

As can be seen, the map consists of a two large clusters (many papers) in the centre, with some smaller clusters in the vicinity and a lot of very small clusters (even fewer papers) scattered around them.[52]

In **Figure 6**, cluster 1 is selected. In the application (and not shown here), this would give this cluster a distinctively darker colour than the other clusters. For this selected cluster, we have retrieved the journals that have been clustered together in the first stage of our map-creation process. The second list contains the first part of the most frequent NPs in the selected journal "Physical Review B". And below this, the clipboard (see next section) is shown, containing the single element "neural network". It also shows the spread of the relative number of papers for the term "neural network". We will get back to that in the next section also.

4.3.3 Noun phrase selection for visualisation

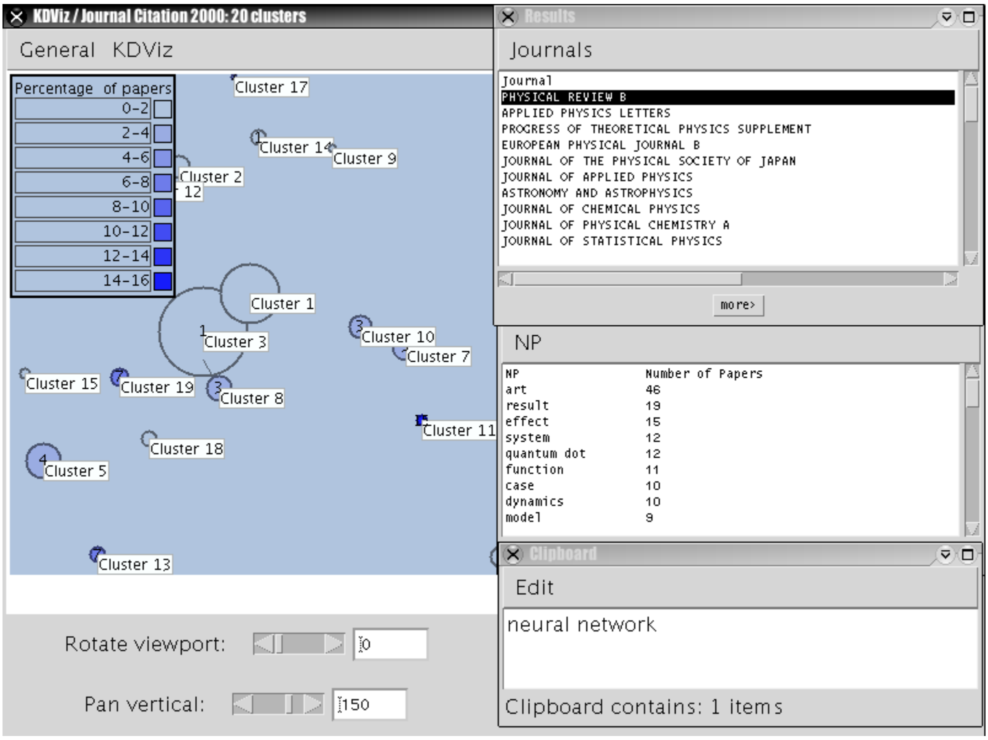
Selecting the NPs for visualisation can currently be done in three ways:

1. from the list of NPs associated with the selected clusters;
2. from the list of NPs associated the journals that have been selected in the list of journals retrieved for the selected clusters; or
3. from the list of NPs that are the result of a search for specific NPs, regardless

---

[51] And some small variants of these phrases, like "complex-system"

[52] The small cluster number 8 (near the centre) has been dragged away from its original position somewhat, as indicated by the small line connecting it to its original position



Journal
PHYSICAL REVIEW B
APPLIED PHYSICS LETTERS
PROGRESS OF THEORETICAL PHYSICS SUPPLEMENT
EUROPEAN PHYSICAL JOURNAL B
JOURNAL OF THE PHYSICAL SOCIETY OF JAPAN
JOURNAL OF APPLIED PHYSICS
ASTRONOMY AND ASTROPHYSICS
JOURNAL OF CHEMICAL PHYSICS
JOURNAL OF PHYSICAL CHEMISTRY A
JOURNAL OF STATISTICAL PHYSICS

NP	Number of Papers
art	46
result	19
effect	15
system	12
quantum dot	12
function	11
case	10
dynamics	10
model	9

neural network

Clipboard contains: 1 items

Rotate viewport: 0

Pan vertical: 150

Figure 6 Overview of the application, with (part of) the list of journals in cluster 1 and (part of) the list of NPs in the selected journal “Physical Review B” and the clipboard containing the single item “neural network”.

of the cluster selection.

The selected NPs are put on a ‘clipboard’. The contents of this clipboard is independent of the selection mechanism, so we may e.g. select a number of journal NPs as well as NPs we specifically searched for.

Figure 6 shows a visualisation of the distribution of the relative number (more on that in the next section) of papers containing the selected NP “neural network”. This is the only element currently on the clipboard. The clusters containing a relatively large number of papers associated with this NP are positioned in like a sickle-like shape from the cluster 13 via cluster 8 to cluster 11. Also, although barely visible, cluster 17 (positioned at the top of the map) has also been given a dark colour. But the term spread is really confined to specific regions of the map only.

4.4 Absolute versus relative

In an earlier version of the implementation it was only possible to display the absolute number of papers associated with the NP selection. But it soon became apparent

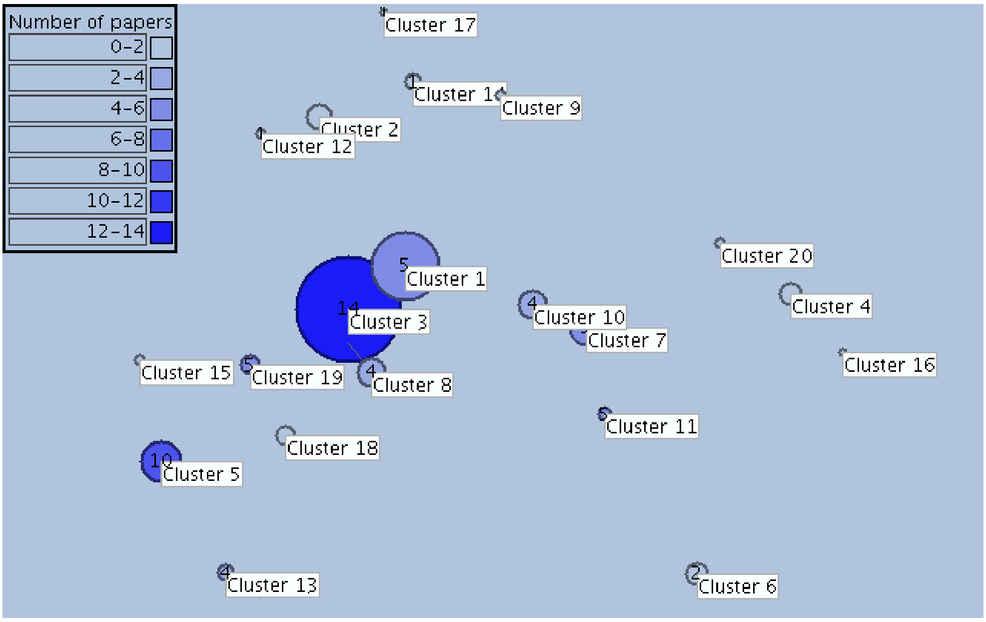


Figure 7 The map showing the spread in absolute numbers of papers for the NP “result”.

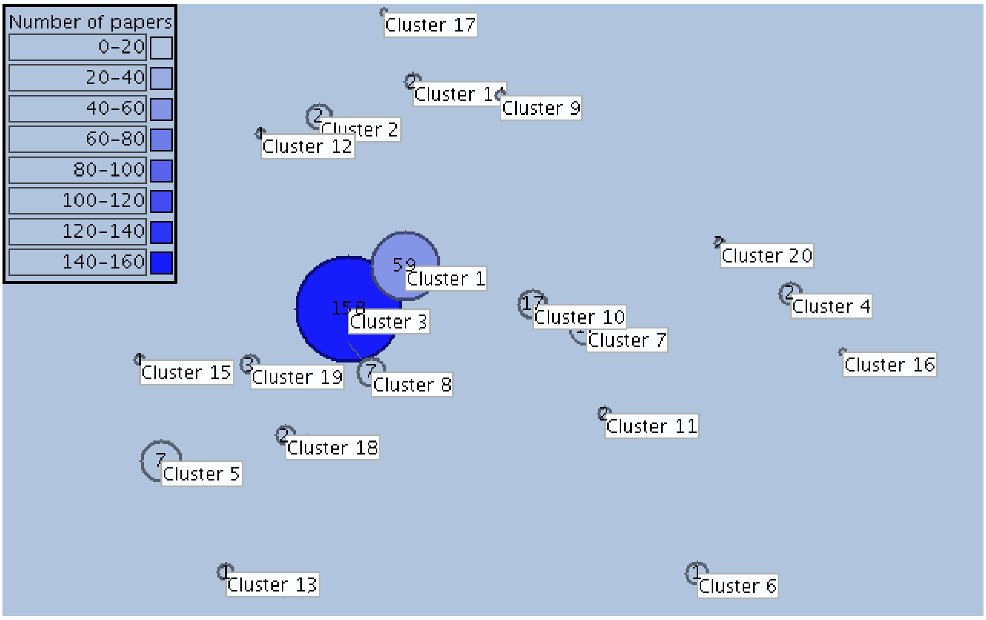


Figure 8 The map showing the spread in absolute numbers of papers for the NP “dynamics”.

that it would be more informational to use the relative number of papers per cluster containing the selected NPs.

The advantage of the use of the relative number of papers as opposed to the absolute number is nicely illustrated by **Figure 7**, **8**, **9**, and **10**. These figures show the spread of the absolute and relative number of papers containing either the NP "result" or "dynamics". Both terms are in the top 10 most frequent NPs in the whole map (shown in **Table 1**) and so they both appear to be generic.

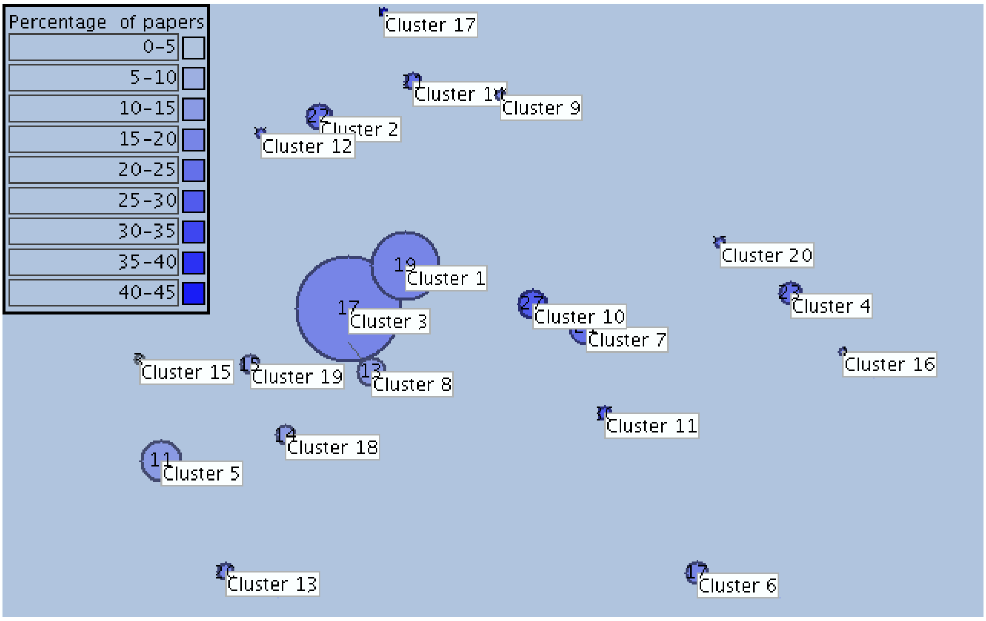


Figure 9 The map showing the spread in relative numbers of papers for the NP "result".

Taking a look at the spread in absolute numbers of papers (**Figure 7** and **Figure 8**), on the one hand confirms our expectation that the terms are generic, since both look rather alike. But on the other hand, it turns out that the structure of the map distorts our visualisation: the larger clusters in the centre of the map contain such a large number of papers, that the smaller clusters all quickly fade into the background. And this is simply because of their lack of papers in general and *not* necessarily due to their lack of papers containing the specific term. As a result, the spread of both terms seems to be confined to the centre of the map, contrary to our expectation that it should more or less cover the whole map.

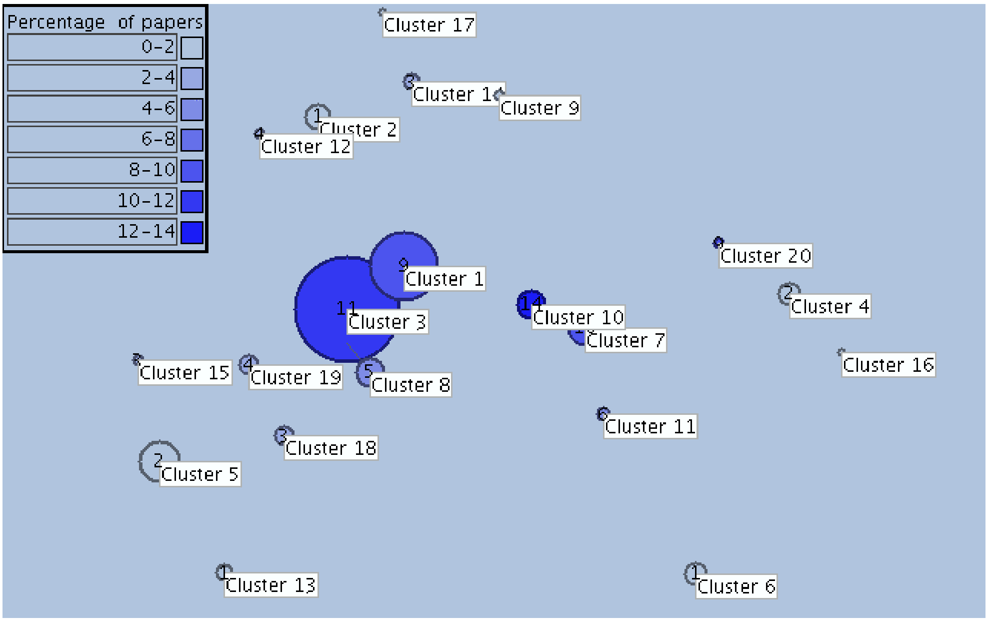


Figure 10 The map showing the spread in relative numbers of papers for the NP "dynamics".

Rank	NP	N
1	c	1220
2	right	748
3	system	656
4	result	608
5	paper	530
6	model	413
7	chaos	381
8	method	348
9	effect	314
10	dynamics	284

Table 1 The 10 most frequent NPs for all the papers in the map used in the examples and the number of distinct papers N.

But if we visualise the distribution of the relative amount of papers over the clusters instead, a different picture emerges. Now, the term "result" indeed proves to be evenly (randomly) spread over the whole of the map. But surprisingly, the term "dynamics" is still confined to the clusters near the centre of the map, which is mainly populated with papers from physics journals. This does not necessarily imply "dynamics" to be content-bearing or usable, but it does require us to take a better look: it could for example be the case that it is used in physics papers in a specialised way if related to "complex systems" or "non-linear dynamics".

4.5 Conclusions

It has proven to be relatively easy to add functionality to our interactive map-interface to visualise the distribution of terms (NPs) as a spread of numbers and colours over a BM. A sample analysis shows that textual elements can be used to relate parts of bibliometric maps, and seems to confirm our assumption that generic terms are spread over the whole map, whereas content-bearing terms are limited to specific regions only. But it was also noted that using the relative number of papers proves to be more informational in a visualisation than the absolute number of papers. This turned out to be especially the case if the number of papers varies substantially between the clusters in the maps, as was the case in our sample analysis.

References

Buter, R. K. and Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68.

Callon, M., Law, J. and Rip, A., editors (1986). *Mapping the Dynamics of Science and Technology.* The MacMillan Press Ltd, London.

Kim, W. and Wilbur, W. J. (2001). Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology, 52*(3), 247–259.

Moed, H., Bruin, R. D. and Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool.* PhD thesis, Leiden: Leiden University.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.

Van Raan, A. F. J. (1999). Advanced bibliometric methods for the evaluation of universities. *Scientometrics, 45*(3), 417–423.

5

Combining concept maps and bibliometric maps: First explorations*

Abstract

Bibliometric maps of science are a well-established research subject. But their adoption as a science policy support tool is lacking. We think this is because the user does not immediately comprehend a map and (as a result) is not enticed into using it. To help this comprehension, we propose the use of 'qualitative maps': an umbrella term for diverse tools such as concept maps and mental maps. We developed a tool that interfaces between a qualitative map and a bibliometric map which lets the user create a correspondence between the distinct vocabularies of the maps. We also conducted two user studies: the first explored the combined use of bibliometric and qualitative maps and the second the preferred format of the map and the word-usage in the description of its elements.

5.1 Introduction

Maps of science have been around for quite a while. They are firmly established as a bibliometric and scientometric research subject, as illustrated by works like (Callon et al., 1986; Braam et al., 1991a; White and McCain, 1998 and Noyons, 1999). Mapping can be considered a bridge to related research in fields such as Knowledge Visualization (Chen, 1999) and Information Retrieval (Lin, 1997).

Bibliometric maps are graphical summaries of sets of papers, either based on citation data, words or phrases or some other bibliometric elements. Part of their appeal is based on the shared characteristics with their geographic counterparts (Small, 1999) allowing them to be referred to as 'landscapes of science' (Noyons, 1999).

Besides the maps deployed in Bibliometrics, a number of (diagrammatic) represen-

---

* Originally appeared in Scientometrics 66(2), 377–387, 2006.
Reproduced with permission of the publisher.

tation schemes has been developed with a distinctive character. These schemes have names such as mental maps, mind maps or concept maps. The appearance of these maps varies: some are like networks, either with directed or undirected relations; some have labelled relations; some are strictly hierarchical; and some are completely free-form. The position of the elements and their relations are laid out manually and are usually from the hand of a single person. In contrast, the structure of bibliometric map is generated algorithmically and supposedly without user bias. We will use the term qualitative map (QM) to refer to these different, but related mapping techniques.

Qualitative maps can, for example, be used to take notes in a lecture or a meeting or lay out knowledge to benefit education (Novak, 1998). And when carefully crafted to represent a cognitive structure of some problem matter, the depictions can also be analysed and used as an inference tool. Examples of matters being analysed this way include: strategies (Huff, 1990) organizational learning (Srinivas and Shekar, 1997) and international politics (Young, 1996).

Even though the research in bibliometric maps has settled itself nicely, the adoption of maps as a policy-supporting tool is lacking. Early problems described by Healey et al. (1986) are partly to blame. But techniques have advanced and problems have been addressed. Yet even recent efforts such as the EU-funded Centres of Excellence project[53] that, among other things, explicitly targeted the further development of bibliometric maps as a science policy support tool have not been able to change this situation.

We have extensive experience with maps in combination with evaluative bibliometric analyses, as illustrated in e.g. Noyons (1999, 2004). Based on this experience, we can say there is not much wrong with the potential of maps. After having been guided through the map, a knowledgeable user is usually quick to see advantages of the use of a bibliometric map (BM). Yet without such guidance, there is less response from users. We could anthropomorphically speak of a "lack of communication" between a BM and its users. And this communication problem is an important reason for the lacking adoption of bibliometric maps as a science policy tool.

As Law (1986) suggests, an author writing a paper should carefully phrase that paper in order to get and keep the attention of a potential reader. The phrasing should be broad enough to attract as many interested readers as possible, yet specific enough to avoid the informed reader to dismiss it as being without value. Maps of science are like scientific papers in this respect: a user should be attracted by the broad overview given by the map and find valuable information in its details.

Since most of our maps are built with textual elements, i.e. with words and phrases taken from papers, we could be using the wrong words to describe (label) the elements in our maps. Then, much like an author rephrasing parts of a paper to increase its readability, we could change the descriptions of particular elements to improve the "communication" between a user and a map. But maps don't have the same verbosity as papers do: only short phrases fit on the limited amount of space.

---

[53] For more information, we refer to the website `http://studies.cwts.nl/projects/ec-coe`.

And variance in vocabulary between users is too large, as described by Furnas et al. (1987), to find a set of labels that fits all. As a result, it will be next to impossible to find a labelling that will be descriptive to most (let alone all) users.

We expect a QM can help here and especially the process of creating one: the process activates the user's vocabulary while the user searches for words to describe the elements, at the same time materializing a user-specific vocabulary in these descriptions. The result is a personal, qualitative description of a problem matter: a depiction of elements relevant to the user, bearing labels that are meaningful to that user and the relations between the elements the user considers important. If we can then map the terms in this description to terms on the bibliometric map, we have created links between the (materialized) internal representation of the user and the bibliometric map.

The research described in this paper explores this idea. But since the research is still in progress, it leaves many technical and fundamental questions untouched. Two main components of our explorations are presented. The first is a tool that acts as an intermediary between a QM and a BM; this tool lets a user create trees of associated terms, with at the root of those trees terms extracted from the qualitative map. The second component is a set of two distinct user studies: the first study investigated how a BM is used in combination with a QM and the second study investigated how the preferred QM of a user may look like.

5.2 Methods and tools

5.2.1 Woven Stories

To give our research a head start, we used an existing implementation of a qualitative mapping tool. This tool was (and still is) developed to explore the utility and applicability of "Woven Stories". The concept of Woven Stories (WS) was originally created to support collaborative writing over the web. But more generally, it can be regarded as a cognitive support tool (Gerdt et al., 2001). It is applicable to different domains, like for example E-Learning or Corporate Strategy planning (Nuutinen et al., 2004).

With WS it is possible to write documents with different authors in a distributed, collaborative manner. This is different from, for example, circulating different versions of the same document between different authors, where one has the problem that only a single author can, at any point in time, be working on the "active" version of the document. WS tries to remedy this by dividing the document into sections to which a user may add comments or which a user can edit. These sections are all stored on a central server, which distributes changes to all participating in ("logged on" to) the story. If the user edits a section, a new version is created and communicated to all other users, allowing different versions of the same section to be written

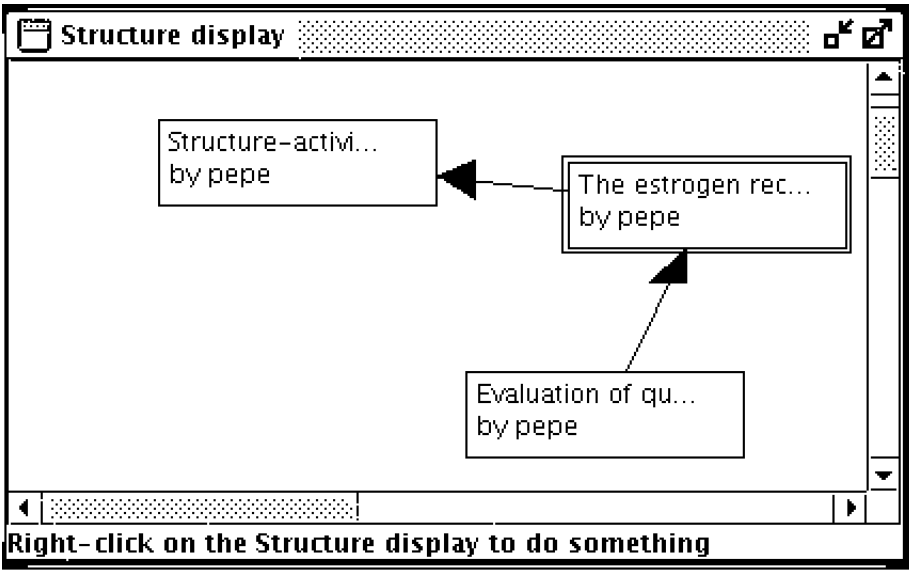


Figure 11 A window from a Woven Stories application, showing three sections (each having a single version) and a single path between these sections. All sections have the fictitious author “pepe”.

parallel.

The different section versions are represented as nodes in a directed graph, as illustrated in **Figure 11**. By following a certain path through the section-graph, a specific version of the document (or “story”) is created. The final document is a specific section path.

We used the sections of WS as the elements of a qualitative map and the paths between those sections as relations between the elements. And, as can be seen from **Figure 11**, the relations are directed and unlabelled. Additional features of the WS tool, such as the distributed nature and the ability to add comments to sections, were not used.

5.2.2 Bibliometric mapping application

The bibliometric mapping application *i*BEX (Interactive Bibliometric EXplorer) we presented in Buter and Noyons (2001) is a mix between an information retrieval and a visualization tool. For example, it can retrieve the list of authors with the highest impact in a certain element or group of elements on a map. But it can also visualize that list by colouring map elements, or display different types of charts either on the elements or on the map as a whole. This application is actively deployed in evaluative bibliometric studies, like the already mentioned Centres of Excellence project. The application had to be slightly extended to be able to communicate with the tool presented below.

5.2.3 Correspondence interface

The choice of words in the descriptions of the QM (its vocabulary) is not necessarily the same as that in the BM. Again, this is because there are many ways a thing or idea can be described. So, multiple related terms are needed to properly describe the research and to retrieve all relevant papers from the database. So, what is needed is a way to map the words or phrases used in a QM to the corresponding ones in the BM.

This correspondence-creating interface, or simply correspondence interface (CI), interactively creates relations between words or phrases from the QM and those of the BM. First, the CI extracts words and phrases from the QM: these are the candidate terms. The user can select some of these terms into a working set, with of course the possibility to extend this working set later on. The correspondence is developed by expanding the terms in the working set into associated terms from the database of the BM. These expansion terms can also be expanded, resulting in an expansion tree.

This is actually something akin to what in Information Retrieval is known as Query Expansion (e.g. Xu and Croft (1996)), where more and more specific queries are created, each one getting closer to retrieve some desired result. However, instead of a query to retrieve documents, a correspondence between the vocabulary used in the QM and the one used in the BM is created. And the result is not a (set of) documents, but a proper coverage of papers (and hence, themes or elements) in the BM.

5.2.4 First user study

The objective of the first user study was to investigate how a QM changes the perception and use of a BM. This was done by employing user scenarios. User scenarios describe the user, the goal of the user, the task the user engages in to reach the goal, the conditions under which the user performs the task and the criteria that determine whether the result (of the task) can be considered a success. All this is formally described in TC 159/SC 4 (2000).

The user scenario we designed for this study, targeted field experts, such as researchers and people involved in science policy. The goal they were given was to search and compare actors considered relevant to a specific research area. To accomplish this, the users had the task to use a BM and to describe the research in a QM, which had to be drawn using some generally available, third-party piece of QM software. For this, the description of the research in the QM must be (considered) in agreement with the research as displayed in the BM. In a successful task, the user confirms the actors are located in the expected parts on the BM.

5.2.5 Second user study

The objective of this study was to investigate the preferred format of a QM and the relation between the word-usage in the BM and the QM. The set-up of this study was rather low-tech: an A3-sized sketchbook and a permanent marker. We did not use a user scenario for this study; instead, we simply instructed the participants, all of them working in the field of Bibliometrics and Scientometrics, to draw a depiction of "bibliometric analysis" on a piece of A3 paper using that permanent marker. To illustrate what a possible result could look like, the participants were shown a depiction of a network of terms. At the same time, the participants were specifically instructed to format their depiction as they saw fit, possibly disregarding the example altogether.

Each participant was given ten minutes to complete the depiction. Somewhere in the last five minutes, the participants were asked to point out their research in the description and to improve the depiction to more accurately represent this research. The two main reasons for these additional questions were to avoid very general descriptions and to direct users who were not sure how to continue.

5.3 Implementation and results

5.3.1 Correspondence Interface

The implementation of a prototype CI resulted in the tool depicted in **Figure 12**. Shown is a working set of five root terms, four of which have been expanded. For example, the tree at the top of the window, sprouting from the root term "xenochemical" has been expanded into three associated terms: "hormone", "milieu" and "contaminant". The last expansion term has also been expanded into two associated ones: "concentration" and "chemical".

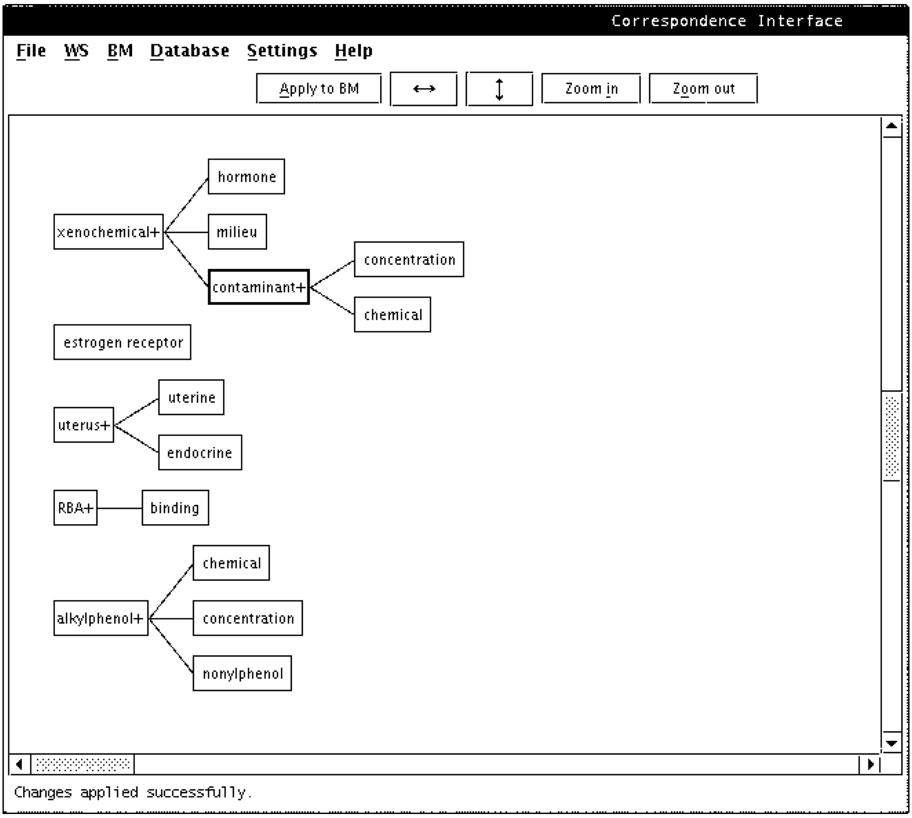
Elements in the CI can be collapsed into a single node, hiding the subtree of that node. To avoid confusion between unexpanded and collapsed elements, an element is given a "+" to indicate it has been expanded and possibly only partly shown.

Expansion terms are selected from a dialog that contains a list of candidate terms, sorted to some criterion. The algorithm that sorts these lists of candidate expansion terms was very straightforward in this prototype: it returned the terms in papers containing the term that is expanded, sorted on frequency of occurrence.

To test the relation between a correspondence and a BM, we implemented two views on the BM:

1. a pie-chart per map-element to show the coverage per selected term in that element;
2. a spread of colours over the map to show the coverage of the complete selection.

In **Figure 13** we show an example of the first type of view: pie-charts on a map of the field "Endocrine Disruptors" for four selected terms. Although "chemical",



Changes applied successfully.

Figure 12 The CI with five root terms, four of which are expanded. The expansion term "contaminant" has also been expanded.

"endocrine" and "concentration" are spread in relatively equal proportions over the map, the occurrence of "contaminant" seems to be less frequently used in papers in the south-east of the map. However, a clear picture does not emerge.

5.3.2 First user study

Unfortunately, it turned out to be hard to find suitable and available experts: only one participant was found, a female researcher in the field of "Endocrine Disruptors". What follows is a description of the execution of the scenario for this participant.

We presented the participant a BM of her field and asked her to assess it. And although there was criticism, the vast majority of the clusters on the map were deemed meaningful in her field.

Then, we asked the participant to create a map that described her field. She divided the research into a limited number of main subjects, which she distributed more or less evenly over the map. To these main subjects she then added associated, related descriptions and drew the relevant inter-subject relations. Notable was the

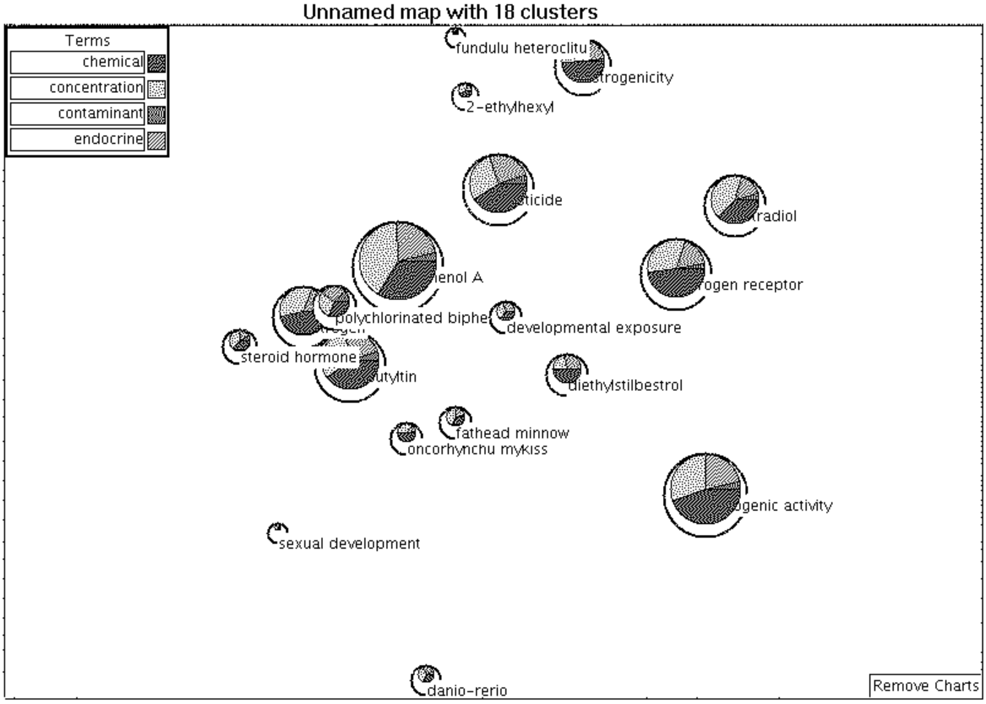


Figure 13 A co-word map in *i*BEX of the field "Endocrine Disruptors" decorated with pie-chart showing the per-element proportional number of papers of four expansion terms.

increased focus on the themes relevant to the research of the participant herself, expressed by adding more details to that part of her map.

After completing the QM, the participant was asked to look at the BM again and describe how it related to her map. All the main subjects from the QM were represented in the map. However, now she found the BM lacking: it was either too detailed or too coarse. As a result, she judged the map not trustworthy enough to continue with the rest of the task.

5.3.3 Second user study

Although there was no generally preferred format, six out of eight participants based their map on a network of descriptions. The other two based it on either a table or a hierarchy. Most participants also grouped descriptions into super-structures, which could have relations to other elements, indicating such relations were applicable to all elements of the group.

The descriptions used in the maps were collected for further analysis. Most descriptions consisted of only one word or phrase, but more complex descriptions were cut into a list of words and phrases (or terms, below). Some terms had to be translated from Dutch to English; the terms that turned out to be untranslatable were discarded.

After translation and spelling unifications (like for example visualisation → visualization) there were 121 distinct terms. The number of distinct terms per user was between 13 and 29, with a median of 19. Of the 121 terms, only 20 (16%) were used more than once. The most frequent term was "bibliometric analysis" (the term given to the participants to describe), appearing in five depictions. The database contained 81 out of the 121 terms. For each participant, between 52 and 93% of the terms could be found in the database, with a median of 74%. Of the terms that could not be found, the majority were not nouns or noun phrases, but for example adjectives, which are not kept readily retrievable in the BM database.

There were 132 different words used in the descriptions. The database contained 122 (92%) of these words. There were 33 words which appeared more than once. Per participant, there were between 19 and 36 distinct words, with a median of 21.50. Per phrase, there were between 0.65 and 1.0 unique words, with a median of 0.78.

We also did a quick analysis of the relations used in the maps. We used only the relations explicitly drawn by the user, not the ones implied by the vicinity of descriptions. A relation was represented by a pair of terms. If a relation was drawn between groups of descriptions, or if a complex description was converted to a list of terms, all possible pairs were used. For example, a relation between a single term and a group of three terms resulted in three distinct pairs. Of course, this inflated the number of pairs considerably, possibly over-representing some terms. For example, in the set of pairs, the most frequently used term was "citation", appearing in ten pairs, but used by only three participants. The result was a list of 121 distinct pairs. The number of pairs per user was between 9 and 25, with a median of 13.5. In these pairs, 90 distinct words were used, 58 appearing more than once, and ten appearing more than five times.

5.4 Discussion and conclusions

We presented our first explorations in the use of concept mapping and related techniques to improve the comprehension and usage of bibliometric maps (BMs). This research is part of our on-going investigation into improving the use of bibliometric maps as a science policy tool. We used the term Qualitative Map (QM) as an umbrella term to refer to the different techniques related to concept mapping.

A supportive observation from the first user study is the increased level of scrutiny after having drawn a qualitative map of the field. This seems to corroborate with the idea that the creative process producing a QM indeed increases focus which allows a more scrutinized judgment of a BM and possibly improving its comprehension. In this study however, this increased scrutiny also invalidated the criterion set for a successful scenario: the user did not trust the BM provided a valid structure of the research.

A probable reason for this failure is the naive field delineation for “Endocrine Disruptors”, which essentially boiled down to papers matching the search “endocr* disr*”. Unfortunately, there was no opportunity to improve this delineation and repeat the exercise. But a proper and complete field definition will always be important. The second user-study showed that a majority of the descriptions from a QM can be found in the database of the BM, but this most likely depends on that field delineation: if complete, the user-vocabulary will be available in the papers they are familiar with or even contributed to. However, if that field-delineation is lacking, terms preferred by the user may very well be missing.

The network format seems to be the most preferred format to use in a BM. Of course, it cannot be ruled out that the example given to the participants, which showed a network of terms, introduced a bias towards the network format. Also interesting is the use of groups (of terms and relations) in the depictions, something not commonly seen in the QM related literature and existing QM tools.

Creating a correspondence between the vocabulary of the QM and that of the BM is required for a proper coverage of the research. The representation of this coverage as a forest of trees has an obvious advantage in allowing multiple senses to be present in different trees in the same correspondence. However, in a specialized research area this is probably rarely needed. And also, if a root term is not found in the database, there is in the current implementation no way of expanding that term. If we changed the representation into a network, a user might be able to relate it to terms already expanded from other terms. Such related terms could than be used as clues for approximate matching.

By letting a user create a QM of research relevant to a given science field, we think the comprehension of a BM can improve and as a result make the BM a more attractive tool. But the kind of problem depicted by the user is not necessarily restricted to the delineation of research. Therefore, a possible future extension could be to ask someone to draw up a description of an interesting current research theme, or of actors and their research focus as perceived by that person and directly decorate such depictions with bibliometric indicators.

Still, for the moment there is still a long way to go before this combined mapping approach is usable for inclusion in the policy supporting tool chain. Nevertheless, we are convinced the ability to materialize qualitative expert knowledge and use it in a bibliometric analysis, even beyond being an assistance in comprehending a BM, has a lot of potential and deserves further attention.

Acknowledgements

We would like to thank An Ghekiere for participating in the first user study. The research described in this paper was conducted as part of the EC-funded research project called “Monitoring Environment and Health R&D in Relation to Socio-Economic Problems: a New Approach to Impact Assessment” (contract number QLK4-CT-2002-30244).

References

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Buter, R. K. and Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68.

Callon, M., Law, J. and Rip, A., editors (1986). *Mapping the Dynamics of Science and Technology*. The MacMillan Press Ltd, London.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management, 35*(3), 401-420.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.

Gerdt, P., Kommers, P., Looi, C. K. and Sutinen, E. (2001). Cognitive technology 2001. In Beynon, M., Nehaniv, C. L. and Dautenhahn, K., editors, *LNAI 2117*, pages 233–247. Springer-Verlag, Berlin.

Healey, P., Rothman, H. and Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233–251.

Huff, A. S. (1990). *Mapping strategic thought*. Chichester: John Wiley & Sons.

Law, J. (1986). The heterogeneity of texts. In Callon, M., Law, J. and Rip, A., editors, *Mapping the dynamics of science and technology*, pages 67–83. London: The MacMillan Press Ltd.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40–54.

Novak, J. D. (1998). *Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations*. New Jersey: Lawrence Erlbaum Associates.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool*. PhD thesis, Leiden: Leiden University.

Noyons, E. C. M. (2004). Science maps within a science policy context. In Moed, H. F., Glänzel, W. and Schmoch, U., editors, *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&t Systems*, pages 237–255. Dordrecht: Kluwer Academic Publishers.

Nuutinen, J., Sutinen, E., Liinamaa, K. and Vanharanta, H. (2004). Woven stories as a tool for corporate strategy planning. In *Proceedings of the WBC2004*, pages 438–441. Lisbon.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.

Srinivas, V. and Shekar, B. (1997). Applications of uncertainty-based mental models in organizational learning: A case study in the indian automobile industry. *Accounting, Management and Information Technologies, 7*(2), 87–112.

TC 159/SC 4 (2000). Human-centred lifecycle process descriptions. *ISO standard 18529*.

White, H. D. and McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11.ACM

Young, M. D. (1996). Cognitive mapping meets semantic networks. *Journal of Conflict Resolution, 40*(3), 395.

6

Non-alphanumeric characters in titles of scientific publications: an analysisof their occurrence and correlation with citation impact*

Abstract

We investigated the occurrence of non-alphanumeric characters in a randomized subset of over almost 650, 000 *titles of scientific publications from the Web of Science database. Additionally, for almost* 500, 000 *of these publications we correlated occurrence with impact, using the field-normalised citation metric CPP/FCSm. We compared occurrence and correlation with impact both at in general and for specific disciplines and took into account the variation within sets by (non-parametrically) bootstrapping the calculation of impact values. We also compared use and impact of individual characters in the thirty fields in which non-alphanumeric characters occur most frequently, by using heatmaps that clustered and reordered fields and characters. We conclude that the use of some non-alphanumeric characters, such as the hyphen and colon, is common in most titles and that not including such characters generally correlates negatively with impact. Specific disciplines on the other hand, may show either a negative, absent, or positive correlation. We also found that thematically related science fields use non-alphanumeric characters in comparable numbers, but that impact associated with such characters shows a less strong thematic relation. Overall, it appears that authors cannot influence success of publications by including non-alphanumeric characters in fields where this is not already commonplace.*

---

* Originally published in Journal of Informetrics 5, pp. 608–617, 2011.
Reproduced with permission of the publisher.

6.1 Introduction

Every day, the inbox of a modern researcher readily fills up with emails from friends, colleagues, and even complete strangers. Even more, at stated intervals, emails arrive that contain titles of interesting publications which have recently been added to databases such as Pubmed, Scopus, or the Web of Science. Furthermore, personal messages, electronic forums, web sites, and social networks all require attention and time. Evidently, new scientific literature is only one stream of information that nowadays flows towards a researcher—albeit a rather pivotal one for the profession at hand. Already some time ago, Meadows (1998) estimated that an average researcher had to scan through roughly $3,000$ titles per year. We assume that this has only become more, and that the increased information burden leaves even less time to deal with them. Clearly, to get attention of potential readers, it is crucial that a publication is presented effectively to a researcher. In many cases, the title is the way to accomplish this. Of course, an author could try a tactic employed by writers of certain emails BEGGING FOR attention. Yet, there is a good chance that this will annoy and subsequently put off potential readers, and since being read is an important factor in the professional success of authors, this is evidently not desirable. As writing and publishing is a communal effort, readers are used to certain topics and styles. Authors can use this to their benefit, by using familiar ways of phrasing a title in order to facilitate quick reading and to use signal words that are expected to trigger the interest of an audience. Yet, phrasing a title too general can bore: a title has to stand out too. Standing out can be accomplished by phrasing differently, for example by using a well-known (but within science not common) literary template such as "to X or not to X" (and filling at the X the particular topic of interest). Alternatively, it could be as simple as using particular, non-alphanumeric characters in a title.

Specific non-alphanumeric characters and title characteristics have been the subject of previous research. Early studies by Dillon (1981, 1982b) showed that the colon (":") has become a standard character in titles of scientific publications. Lewison and Hartley (2005) also studied the colon and found differences in title length and colon usage, both over time and over disciplines. Hartley (2007) combined a meta-analysis with new results and showed that colons are preferred by students because they improve the structure of a title, is not necessarily appreciated by their fellow academics, who make up the intended audience of most scientific publications. However, studies cited by Hartley (2007) failed to find significant differences between the number of citations for publications with and without colons in their title, although the scope of this result was limited to a single journal. Beside the colon, Ball (2009) showed that the question mark has become a frequently appearing in titles in Medicine and (to a lesser extend) in Physics. We generalize these previous studies on specific aspects of titles and investigate both use of specific characters in publication titles and correlation with impact in a broad and extensive sense. By this, we mean that we do not focus on a particular (non-alphanumeric) character nor limit our investigation to specific journals or science fields.

Our main research question is: given the importance of readership in the success of scientific publications, could something simple as using a particular type of character "boost" the success of a publication. Our hypothesis is that the effect of non-alphanumeric characters on the success of publications is constrained by conventions regarding readability and form. Consequently, if such characters occur and exhibit a positive correlation with the success of publications, those characters usually have a known function or are accepted elements. We investigate this by posing the following research questions. First, what non-alphanumeric characters exist in scientific publications? Then, can we see a difference in the success of publication with and without such characters? Also, are such effects global, or can we see differences over disciplines? Additionally, what is the effect of frequently occurring characters? Finally, how does the use and impact of characters compare over fields?

6.2 Method

To investigate non-alphanumeric characters in titles, we extracted publications from all research fields available in the Web of Science database[54] (WoS) published in the period 1999-2008. However, the number of publications available in the WoS for that period is large (almost 13 million), which makes exhaustive analyses too time-consuming and we therefore took a representative, five percent random sample from the WoS, reducing the number of publications to almost 650,000.

To extract the non-alphanumeric characters from the titles of the publications in this set, we used simple regular expression[55] (Aho, 1990). By matching titles with this expression, we got for every publication a (possibly empty) list of non-alphanumeric characters. If this list was empty, we regarded a title as "alphanumeric", and "non-alphanumeric" otherwise.

To express the success of a publication we chose citation counts. Obviously, success expressed in citations is not the same as success expressed in readers (Moed, 2005a). Still, we consider this metric appropriate in the context of scientific success, as well as (not unimportantly) generally easier to obtain than number of readers. To calculate the citation rates we proceeded as follows. For the (almost 500,000) articles, letters, notes, and reviews in our sample, we counted the citations from the (almost 13 million) other publications in the WoS published in the same period. Unfortunately, absolute citation counts cannot be used to compare publications published in different fields, because the (average) length of a reference list differs from field to field and hence the expected number of citations. Therefore, we used the CWTS CPP/FCSm

---

[54] This WoS database is available to the CWTS under license from its publisher Thomson Scientific and contains publications published from 1980 onwards.

[55] We used the Perl-like regular expression `[^\w\s]` which has the following meaning: the square brackets indicate a character set consisting of the character families `\w` (all alphanumeric characters) and `\s` (all whitespace characters), which is negated by the caret (`^`); so this expresses a match of characters which are neither alphanumeric nor whitespace.

indicator, a field-normalised citation metric that employs the WoS journal subject categories (JSCs) as proxies for fields (Moed et al., 1995 and Van Raan, 1996). By using this metric, we can compare citation counts over fields. We acknowledge that JSCs have well-known problems with respect to the delineation of related research, and that they provide a rather high-level classification with only 243 categories to represent all scientific research. Nevertheless, they are a readily available categorisation of related publications, with a definition that is standardized in a transparent way and quite stable over time. We measured correlation between the use of non-alphanumeric characters and impact both in general (whole of science) and at the discipline level (such as "mathematics" or "clinical medicine"). Because we regard disciplines as aggregations of related science fields, we also used aggregations of JSCs to represent disciplines. These aggregations have been developed at CWTS (Tijssen et al., 2008) and are actively maintained[56].

In every set of publications, the value of a metric such as the CPP/FCSm differs from paper to paper. To compare values associated with particular sets of publications, we should take into account this variation when we judge how strong the difference between observed values is. We used bootstrapping (Efron and Tibshirani, 1993) to accomplish this. Bootstrapping is a technique which repeatedly resamples observations with replication[57], and calculates a statistic for every resampled set[58]. The collected values of these calculations can be regarded as an approximation of the distribution around some mean. The use of bootstrapping for this purpose is not new. In the past, bootstrapping has for example been used by Barber and Thompson (2000) to analyse cost data related to patient care, while taking into account the skewed distribution of such data. In addition, Chaudhary and Stearns (1996) used it to analyse the cost-effectiveness of trials. In our study, we recalculated the CPP/FCSm for 10,000 samples[59]. The result we used to create density plots: if we saw the plots hardly overlapped, then we assumed that this indicated that impact values associated with these sets were different.

Some non-alphanumeric characters are quite common tools in English words and phrases. To assess the effect of those common non-alphanumeric characters, we repeat our analyses twice: once including all non-alphanumeric characters and once without the five most frequent non-alphanumeric characters. Because the second analysis excludes titles that only have frequently occurring non-alphanumeric characters, we get an impression of the effect of less common non-alphanumeric characters in titles.

Finally, to compare individual non-alphanumeric characters in fields in which they

---

[56] Maintenance is needed because the JSCs are not stable, but gradually change over time.

[57] Which means that observations can occur multiple times in a sample, because each time an observation is chosen from the set of observations, it is not removed from that set, so it can be chosen again.

[58] A more detailed description on how and when to apply bootstrapping techniques can be found in Carpenter and Bithell (2000).

[59] This number is arbitrary, yet well beyond the 1,000 to 2,000 repeats suggested in Efron and Tibshirani (1993). Current computer power makes such large repeats possible; however, even larger repeats are not useful, as the do not significantly change the observed characteristics of the distribution of

occurred frequently, we employed a matrix visualisation called a "heatmap" in the statistical environment R (R, 2005). This visualisation is a matrix of coloured cells and two dendrograms: one at the top and one at the right of the matrix. These dendrograms are the result of separately clustering the rows and columns of the matrix, and are used to reorder these rows and columns. The heatmap allowed us to assess two different (but related) issues at the same time: fields that appeared close together showed similar use or impact of characters; while characters that appeared close together, showed similar effect over fields. For example, if Hematology and Oncology appeared as two subsequent rows in the heatmap for occurrence, this meant that their use of certain characters in titles is similar. We coloured cells using a gradient of grey colours, with white to indicate nearly zero and black to indicate a value near the maximum. To cluster, we employed the (in R) default hierarchical clustering algorithm (complete linkage) and Euclidean distance.

6.3 Results and Discussion

6.3.1 Occurrence of non-alphanumeric characters

Our five percent random sample consisted of 642,807 WoS publications, all published between 1999 and 2008. **Table 2** lists the 29 non-alphanumeric characters we encountered in the titles of these publications. Next to rank (#) and character (C), this table also shows the number of publications (N) associated with a character, as well as the percentage (%) relative to all publications (in the sample); a point estimate of the impact (I); and the number of publications (articles, letters, notes, reviews) used to calculate that impact (N2). In total, 434,072 publications (68%) have at least one non-alphanumeric character in their title, most frequently a (combination of) hyphen, colon, comma or parenthesis. Overall, we counted 1,102,904 occurrences of non-alphanumeric characters, which is about two percent of the 56,148,591 total character occurrences.

Interestingly, **Table 2** shows a slight difference between the occurrences of characters which ought to appear in pairs, such as "(" and ")", or "{" and "}". After looking up some original copies of publications which showed this asymmetry, we conclude that these differences are database-processing errors.

6.3.2 Occurrence over time

Some authors have suggested that the occurrence of specific non-alphanumeric characters in titles, like the colon (Dillon, 1982b) and the question mark (Ball, 2009), increases over time. It is therefore interesting to see if this is also the case for the complete ensemble of non-alphanumeric characters in **Table 2**. To accomplish this, we counted the number of publications per year with and without at least one non-

#	C	N	%	I	N2
1	-	266,506	41.5	1.01	210,695
2	:	125,316	19.5	1.1	167,816
3	,	83,960	13.1	0.99	58,704
4	)	69,849	10.9	0.89	50,334
5	(	69,846	10.9	0.89	50,335
6	.	57,010	8.9	0.81	19,981
7	’	35,029	5.5	0.83	19,319
8	/	19,244	3.0	1.1	15,446
8	?	18,102	2.8	1.02	11,747
9	+	6,462	1.0	0.96	5,355
10	"	6,303	1.0	0.86	3,901
11	[	4,612	0.7	0.73	4,210
12	]	4,591	0.7	0.73	4,193
13	=	2,256	0.4	0.91	2,168
14	;	1,577	0.3	0.78	1,036
15	&	1,328	0.2	0.49	631
16	>	835	0.1	1.09	731
17	!	736	0.1	0.44	354
18	%	613	0.1	1.15	468
19	*	521	0.1	0.91	439
20	{	438	0.1	0.91	423
21	}	425	0.1	0.91	410
22	<	377	0.1	0.95	322
23	@	89	0.0	1.14	75
24	$	64	0.0	0.21	20
25	#	56	0.0	0.82	34
26	_	45	0.0	0.34	32
27	\	14	0.0	0.50	13
28	|	3	0.0	0.30	3
29	‘	1	0.0	0	0

Table 2 The 29 non-alphanumeric characters ranked by number of publications N; also given is the occurrence as a percentage (%) relative to the total number (642,807) of publications in the set, a point-estimate of the impact I, and the number of publications N2 (article, letters, notes, and reviews) used to calculate the impact.

alphanumeric character in their title, as well as the percentage relative to the total number of publications in that year (in the sample). The result is **Table 2**, which shows that the absolute number of publications with one or more non-alphanumeric characters increases steadily over time, while the relative number of such titles in the sample is stable. This suggests that there is no general change or trend in the use of these characters in the last decade. As such, our finding contrasts those earlier

Year	N	N_{With}	%
1999	56,441	38,131	68
2000	55,929	38,079	68
2001	57,855	39,566	68
2002	56,178	38,140	68
2003	62,738	42,019	67
2004	60,212	40,831	68
2005	73,547	49,647	68
2006	68,292	45,927	67
2007	70,663	47,955	68
2008	80,952	53,777	66
Total	642,807	434,072	68

Table 3 The number of publications N in the sample, and the N_{With} that have at least one non-alphanumeric character, as well as the percentage of publications with such a character.

results.

6.3.3 Examples of titles with non-alphanumeric characters

Table 2 and **Table 3** clearly show that some non-alphanumeric characters appear frequently in titles of scientific publications. Even more, most publications have at least one non-alphanumeric character. Still, the occurrence of characters such as the "@", "|", and "*" can be regarded as more "surprising". To provide more context of their use, we show some examples[60]. Our first example is the use of "@" in the following chemistry-related title:

> Synthesis of Au@SiO2 Core/Shell Nanoparticles and their Dispersion into an Acrylic Photocurable Formulation: Film Preparation and Characterization (Sangermano et al., 2008)

The next title shows a similar use of the "@", as well as a particular use of the dot and the hyphen:

> Design of Rh@Ce0.2Zr0.8O2-Al2O3 nanocomposite for ethanol steam reforming (Rogatis et al., 2008)

These two examples show that the "@" is functional, because it has meaning attached: apparently it expresses (nano-) particles (Au, Rh) attached to some surface (such as SiO2). However, the following use of "@" in a genomics paper has no meaning and is clearly decorative:

---

[60] The examples are extracted from the whole WoS and not exclusively from our sample.

> ISOL@: an Italian SOLAnaceae genomics resource (Chiusano et al., 2008).

Chemistry harbours more examples of the functional use of non-alphanumeric characters. For example there appears to be a convention to indicate “microinterfaces” with a vertical bar “|”:

> Voltammetry of tetraalkylammonium picrates at water|nitrobenzene and water|dichloroethane microinterfaces; influence of partition phenomena (Sladkov et al., 2004)

Outside of chemistry, there are also examples of the functional use of non-alphanumeric characters, although admittedly less spectacular and with a meaning that is less clear:

> *Bhrather “Brother” + Indoeuropean Etymology (Parvulescu,)

However, the humanities too have fine examples of purely decorative use:

> ***!!!The Battle Of The Century!!!*** (Goldbarth, 1997)

We finish with an example that combines an impressive number of distinct non-alphanumeric characters in a single title, all used functionally:

> Endohedral (X@ZniSi)i=4-160,± Nanoclusters, X = Li, Na, K, Cl, Br (Matxain et al., 2007)

6.3.4 Correlation between use and impact

To measure the correlation between impact and the use of non-alphanumeric characters, we bootstrapped the calculation of the impact indicator (CPP/FCSm) for the 482,637 (75% of the 642,807 publications in our set) articles, letters, notes, and reviews in our sample.

6.3.4.1 Impact differences in the whole sample

First, we compared the impact of publications with and without non-alphanumeric characters, as well as the impact of the whole sample. The left plot in **Figure 14** shows the result. In this plot, and in all subsequent similar ones, we use a striped line for the group of non-alphanumeric publications, a continuous line for the alphanumeric publications, and a dotted line for the whole sample. The plot shows a difference of about 0.10 between the alphanumeric and the non-alphanumeric point estimates of impact. In addition, judging from the density plots around these estimates, this difference is significant. The difference between the non-alphanumeric set and the overall result is much smaller, and their plots show considerable overlap. As the majority of papers in the sample has at least one non-alphanumeric character (as became apparent from **Table 2**), this overlap is not surprising.

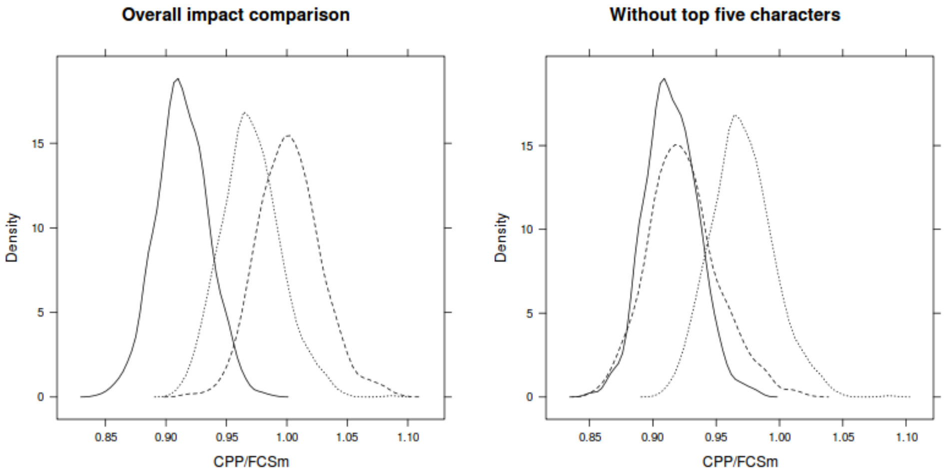


Figure 14 Density (estimated using bootstrapping) plots for the distribution of the CPP/FCSm value in publications with at least one non-alphanumeric character (striped), only alphanumeric characters (solid), as well as the overall impact (dotted). The left plot shows the overall comparison, and the right plot shows the comparison which excludes the top five most frequent characters (see **Table 2**) from the non-alphanumeric set.

The result of the repeated analysis disregarding the top five most frequent characters is shown in the right plot of **Figure 14**. The plot shows that the non-alphanumeric set now has an impact which overlaps considerably with the alphanumeric set. Apparently, the larger impact of publications in the non-alphanumeric set in the left plot is mostly associated with publications that have titles with frequently occurring non-alphanumeric characters. Column I of **Table 2** supports this conclusion, since values become smaller (below 1.0) for less frequently occurring characters. A notable exception is the slash "/" (1.1 for over 15,000 publications). Also note that frequent characters such as "(", ")", and ".", correlate negatively with impact at this global level.

6.3.4.2 Number of non-alphanumeric characters and impact

Given that the occurrence of at least one non-alphanumeric character correlates positively with impact, we wondered if the occurrence of more than one showed additional (positive or negative) influence. To assess this, we created the plot in **Figure 15**, which shows impact values for publications with a given number of non-alphanumeric characters. The result also shows a linear estimation of the trend, which is almost horizontal and we thus see no evidence for a significant positive (or negative) correlation. At most, we can conclude that publications with a larger number of non-alphanumeric characters show a smaller variation in impact; but this is due to the increasingly smaller number of publications.

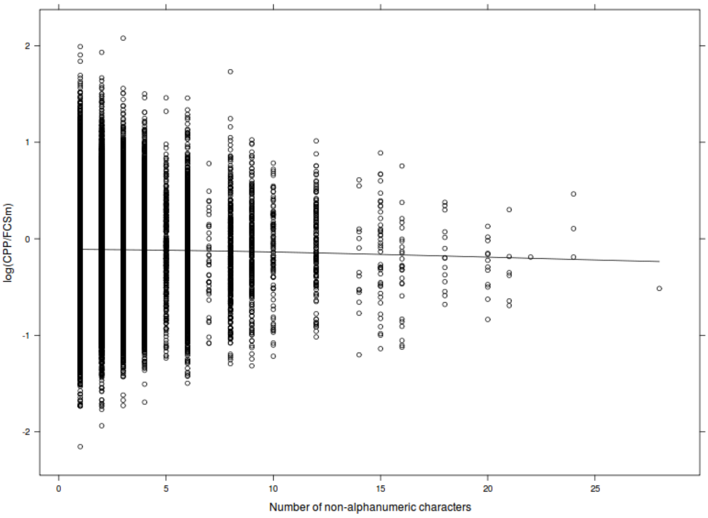


Figure 15 The correlation between the impact (given as the logarithm of the CPP/FCSm value) and the number of non-alphanumeric characters in a single title, together with a linear estimation of the trend.

6.3.4.3 Examples of impact in three disciplines

We finish our impact analysis with three selected examples of impact correlations at the level of disciplines. We chose these three (Clinical Medicine, Agriculture and Food Science, and Biological Sciences) because each one illustrates a different correlation between impact and the use of non-alphanumeric characters in titles.

Figure 16 shows two rows of impact analyses: at the top using all characters and at the bottom without the top five non-alphanumeric characters. We first compare the plots on the top row. The left one, for Clinical Medicines, shows a higher impact of non-alphanumeric publications, similar to the left plot in **Figure 14**. The middle plot for Agriculture and Food Science on the other hand, shows no significant difference between the groups. The right plot for Biological Sciences finally, shows a reversed effect where the alphanumeric titles have a larger impact. These plots show that individual disciplines can exhibit other correlations than at the global level. Still, the bottom row makes this even clearer. Again, for Clinical Medicine, we see a result similar to the right plot in **Figure 14**, where the positive effect of non-alphanumeric characters has disappeared. However, the plot for Agriculture and Food Science publications now shows a positive effect for non-alphanumeric publications, where none is visible in the plot above it. Finally, the plot for Biological Sciences shows an amplified difference of the plot above it, in which the non-alphanumeric set has even less impact.

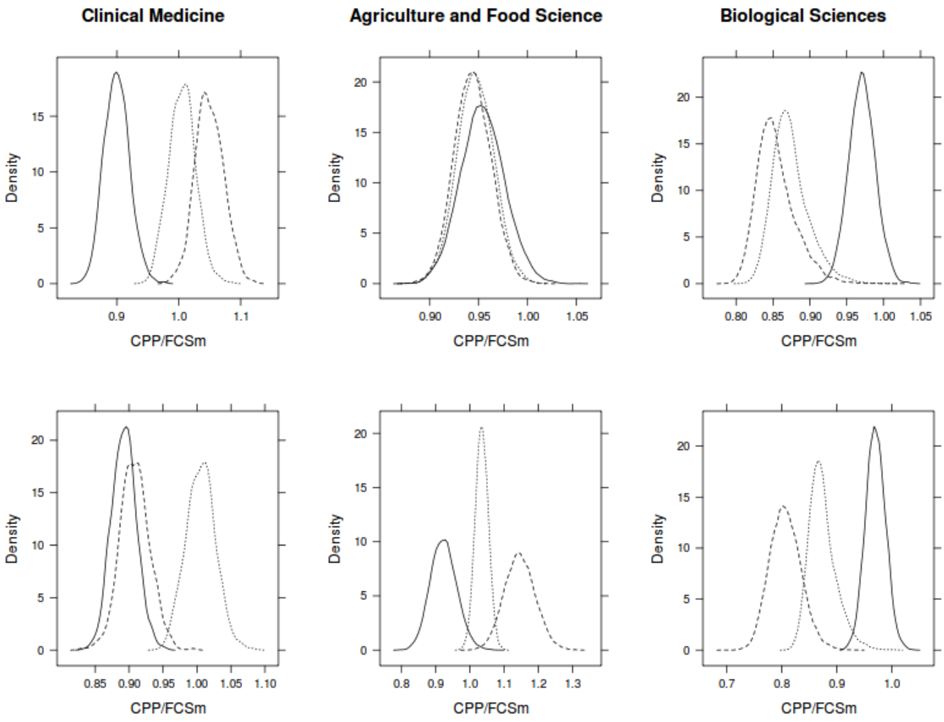


Figure 16 Density plots for the distribution of the CPP/FCSm value for three selected major science fields, for the publications with at least one non-alphanumeric character (striped), only alphanumeric characters (solid), as well as the overall impact (dotted), estimated using bootstrapping. The first row shows the overall results, while the bottom row shows those in which the top five characters are excluded from the non-alphanumeric range.

6.3.5 Fields and characters

We now move away from the global and discipline levels and continue our analyses at the level of science fields and individual characters. First, we show some general field-specific numbers, after which we investigate the use and impact of individual characters in and over fields.

6.3.5.1 Occurrence in fields

Table 4 lists the ten fields which have the most (in absolute numbers) publications containing at least one non-alphanumeric character (column N). The first two fields are chemistry-related, followed by three medical fields. In general, the physical sciences and the life sciences appear to be equally present. Nevertheless, we also note that most life science fields have a clear biochemical association.

#	Field	N	N2	%
1	Biochemistry and Molecular Biology	22183	5075	23%
2	Chemistry, Multidisciplinary	14872	7132	48%
3	Neurosciences	12491	3045	24%
4	Oncology	12336	3645	30%
5	Pharmacology and Pharmacy	12157	2820	23%
6	Materials Science (Multidisciplinary)	11678	2654	23%
7	Chemistry (Physical)	11582	3171	27%
8	Cell Biology	11291	3102	27%
9	Surgery	10678	2544	24%
10	Physics (Applied)	10576	2450	23%

Table 4 The ten fields with the most publications N containing a non-alphanumeric character, as well as the number of publications N2 with at least one non-alphanumeric character outside the top five most frequently occurring characters (see **Table 2**) and the ratio N2/N expressed as a percentage (%).

Additionally, in column N2 the number of publications is given that include a non-alphanumeric character outside the top five in their title. For most fields, the number of occurrences of non-alphanumeric characters is about 25 percent. Notable exceptions are Chemistry, Multidisciplinary and (to a lesser degree) Oncology. The high number of Chemistry, Multidisciplinary correlates with the examples we showed above.

6.3.5.2 Use and impact of characters over fields

Table 4 and the example titles already showed that some non-alphanumeric characters are used in field-specific ways. To further investigate this, we created a heatmap of the top 30 (ordered by descending N) fields and the 29 non-alphanumeric characters. **Figure 17** shows a map where a cell for field F and character C is coloured according to the number of publications that have at least one such character in their title. The clustering of characters visualised by the dendrogram at the top, shows that characters appear divided in four differently graded “bands” of occurrence. The clustering of the fields visualised by the dendrogram at the left, shows that thematically related fields show a similar use of certain non-alphanumeric characters.

Overall, two main branches exist in the clustering of fields: one which starts at the first row down to the row History, and a second one below that. The first branch contains mostly medical fields mixed with some biology and other fields, and the second branch only contains fields from chemistry and physics.

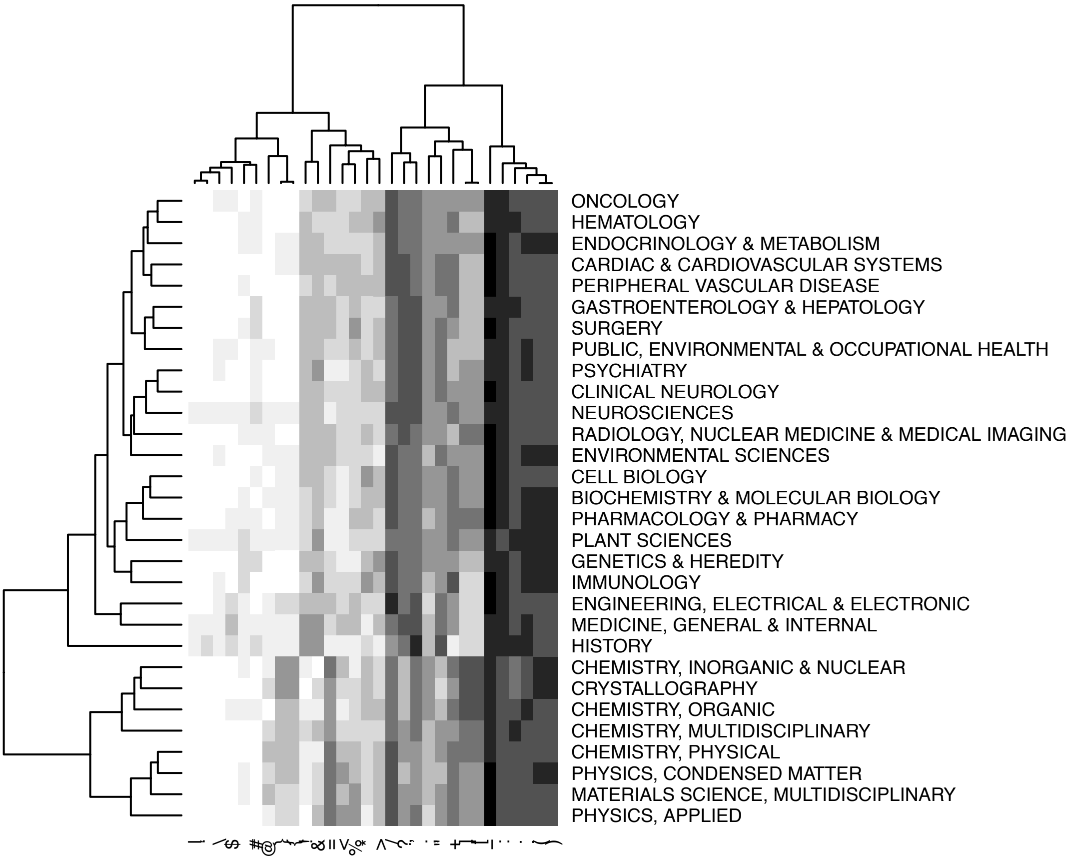


Figure 17 A heatmap of the occurrence matrix of non-alphanumeric characters per field: the dendrogram at the top of the map shows a clustering of the characters shown at the bottom; the dendrogram at the left of the map shows a clustering of the fields shown at the right; darker colours indicate higher (normalised) frequency of occurrence of a character in a field.

We created a second heatmap using the same fields as in **Figure 17**, but now a cell is coloured as a function of impact instead of occurrence. The resulting **Figure 18** shows that impact does not divide characters and fields in such clear groups as frequency did in **Figure 17**. For example, we do not see “bands” of high-impact like the one we found in **Figure 17** for frequency, but values are more dispersed. Still, we can observe some branches of related fields: a high-level chemistry group as well as a medical group. Additionally, Cardiac & Cardiovascular Systems is now an outlier and placed in a branch with History; but in **Figure 17**, it is in the branch with other medical fields.

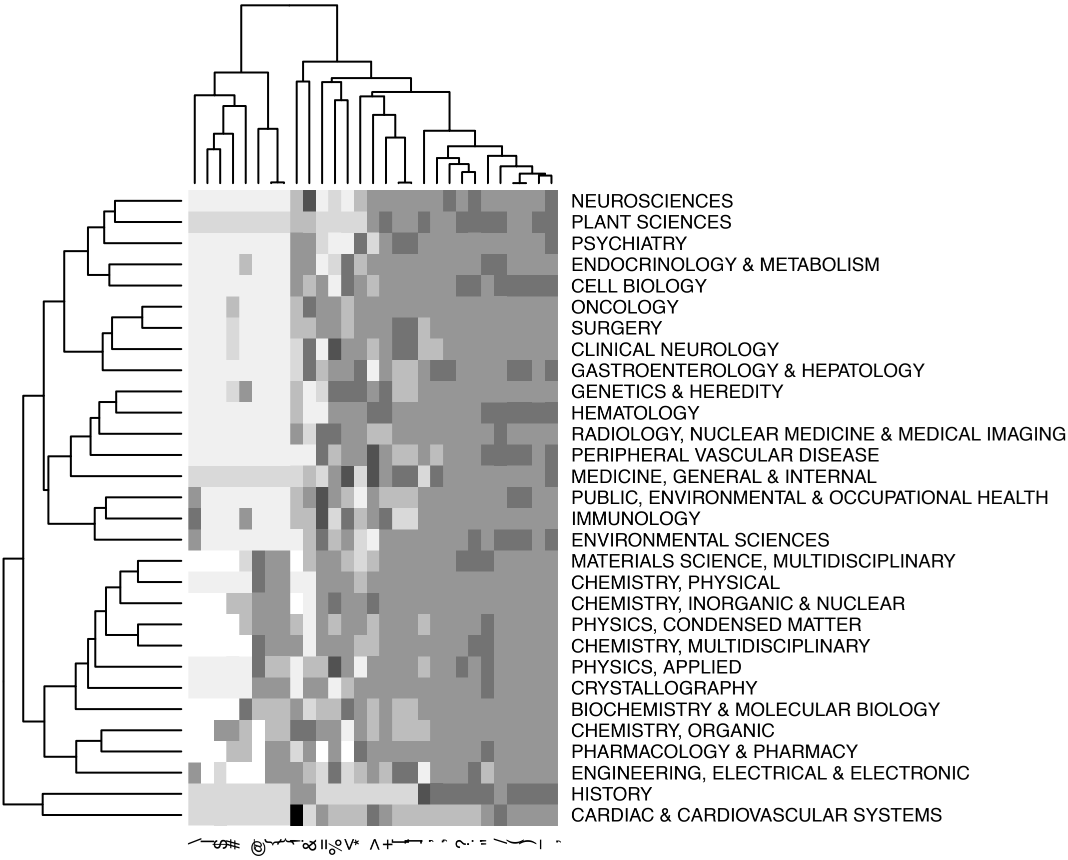


Figure 18 A heatmap like in **Figure 17**, but now using the CPP/FCSm value of non-alphanumeric characters per field, instead of the frequency of occurrence of a character in a field

6.4 Conclusions

We started this publication by pointing out searching for potentially interesting scientific publications has become only one of many sources of information that require the attention of a modern researcher. We then continued to hypothesize that the title of a publication which wants to capture an audience needs to strike a balance between conforming and surprising. An author has several options to accomplish this, for example by sing characters that are neither letters nor numbers. The use of such characters and the correlation with success of publications that use them is the main topic of our paper.

We found that in the representative random sample of publications from the Web of Science (WoS) database, 29 different non-alphanumeric characters occur. Of course, this specific set of characters could be a database artefact, where the database producer has replaced characters that are even more "surprising" by approximations or left them out completely. Additionally, our analyses make clear that using non-alphanumeric characters in a title can be considered normal for an average publication in the WoS, while not doing so is less normal. We also found that the relative occurrence of these characters in titles is stable. This finding contrasts earlier work, which focused on absolute numbers and suggested an increasing use. Selected examples showed that some characters are used functionally and may have meaning associated with them, such as the "@". Most of these functional uses came from chemistry.

In general, publications with titles that have non-alphanumeric characters have a larger impact than publications without. Although the difference is small (0.10), the density plots we created using bootstrapping nevertheless showed it to be significant. We also calculated impact without the top five most frequently occurring characters, and we then saw that the remaining non-alphanumeric publications exhibited lower impact values. However, we showed examples at the level of disciplines in which the correlation is absent or even reversed. In addition, an increasing number of non-alphanumeric characters in a single title showed no additional effects on impact.

Furthermore, we compared fields with respect to the use and impact of non-alphanumeric characters. As already apparent from the examples, fields from chemistry are shown to be an important source of titles containing non-alphanumeric characters, together with life science fields with a biochemical element associated with them. The heatmaps we created show that fields cluster strongly on use, but less on impact. The interesting aspect of this finding is that, although these fields are topically related, this thematic relation is replicated through similar use or (to a lesser degree) impact of single characters. Again, this indicates that the use (and acceptance of using) such characters is shared over related fields.

We think that our observations fit our hypothesis that the most successful publications should strike a balance between convention and surprise. For example, not including any non-alphanumeric characters is just as "odd" as including more exotic characters, unless such exotic characters have a specific meaning (as shown by the chemistry examples). An author therefore cannot artificially boost the success of a publication by including arbitrary non-alphanumeric characters. This is in line with Sagi and Yechiam (2008), who found that humour is generally not appreciated (and thus uncommon) as a mode of discourse in scientific titles; and also with Hartley (2007), who demonstrated the reluctant acceptance of the colon as a standard character in titles of scientific publications. To put things in a broader perspective Haslam et al. (2008) already noted that besides the structure and language of a title, there are many additional factors that influence the prestige of an article: author eminence, journal prestige, and the length of a publication were all shown to correlate strongly with the number of citations. Future research will also move beyond the analysis of single tokens to the analysis of templates or syntax. One topic could be the placement of the non-alphanumeric characters, as they do not appear at random positions in titles, but are part the phrasing or syntax. Additionally, we consider using stylometrics to measure the complexity of the language used in titles and correlate that with impact. A final idea is to create a dictionary of the function or pragmatics of specific characters, like the "@" and "|" we found in the examples given above.

Such additional analyses would slowly increase our knowledge on what makes a publication get the attention of reader. Because, in the end, that is what it was all about in the first place.

Acknowledgements

We kindly thank the anonymous reviewers for sharpening our arguments, as well as their suggestions for future research.

References

Aho, A. V. (1990). Algorithms for finding patterns in strings. In *Handbook of theoretical computer science (vol. A): algorithms and complexity*, pages 255–300. MIT Press, Cambridge, MA, USA.

Ball, R. (2009). Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966–2005. *Scientometrics, 79*(3), 667–679.

Barber, J. A. and Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine, 19*(23), 3219–3236.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141–1164.

Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine, 15*(13), 1447–1458.

Chiusano, M. L., D'Agostino, N., Traini, A., Licciardello, C., Raimondo, E., Aversano, M., Frusciante, L. and Monti, L. (2008). ISOL@: an italian SOLAnaceae genomics resource. *BMC Bioinformatics, 9*(Suppl 2), S7.

Dillon, J. T. (1981). The emergence of the colon: An empirical correlate of scholarship.. *American Psychologist, 36*(8), 879–884.

Dillon, J. T. (1982b). In pursuit of the colon: A century of scholarly progress: 1880-1980. *The Journal of Higher Education, 53*(1), 93–99.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Goldbarth, A. (1997). ***!!!The battle of the century!!!***. *Kenyon Review, 19*(3-4), 143–147.

Hartley, J. (2007). Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research, 29*(4), 553–568.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J. and Wilson, S. (2008). What makes an article influential? predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.

Lewison, G. and Hartley, J. (2005). What's in a title? numbers of words and the presence of colons. *Scientometrics, 63*(2), 341–356.

Matxain, J. M., Eriksson, L. A., Formoso, E., Piris, M. and Ugalde, J. M. (2007). Endohedral (X@ZniSi)i=4-160,+- nanoclusters, x = li, na, k, cl, br. *The Journal of Physical Chemistry C, 111*(9), 3560–3565.

Meadows, D. (1998). Indicators and information systems for sustainable development. *A report to the Balaton Group*, pp. 1–25.

Moed, H., Bruin, R. D. and Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Moed, H. F. (2005a). *Citation Analysis in Research Evaluation*, volume 9 of *Information Science and Knowledge Management*. Dordrecht: Springer.

Parvulescu, A..*brather "brother" + indoeuropean etymology. *Indogemanische Forschungen, 101*, 99–106.

R (2005). Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

Rogatis, L. D., Montini, T., Casula, M. F. and Fornasiero, P. (2008). Design of Rh@Ce0.2Zr0.8O2–Al2O3 nanocomposite for ethanol steam reforming. *Journal of Alloys and Compounds, 451*, 516–520.

Sagi, I. and Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34*(5), 687, 680.

Sangermano, M., Perruchas, S., Gacoin, T. and Rizza, G. (2008). Synthesis of Au@SiO2 Core/Shell nanoparticles and their dispersion into an acrylic photocurable formulation: Film preparation and characterization. *Macromolecular Chemistry and Physics, 209*(22), 2343–2348.

Sladkov, V., Guillou, V., Peulon, S. and L'Her, M. (2004). Voltammetry of tetraalkylammonium picrates at waternitrobenzene and waterdichloroethane microinterfaces; influence of partition phenomena. *Journal of Electroanalytical Chemistry, 573*(1), 129–138.

Tijssen, R. J. W., Hollanders, H., Van Leeuwen, T. N. and Nederhof, A. J. (2008). Science and Technology Indicators 2008. Netherlands Observatory of Science and Technology. Technical Report Leiden: NOWT.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.

7

Identification of Converging Research Areas using Publication and Citation Data*

Abstract

*Converging research is the emergence of new interdisciplinary research from fields which showed limited mutual interdisciplinary connections before. This paper describes three search strategies to identify converging research using publication and citation data extracted from the Web of Science database, including the Social Sciences and Humanities. The search strategies differ in how they represent the originating research fields and the interdisciplinary connections. The field-to-field references (*FFR*) approach uses journal subject categories (*JSC*s), and citations from one category to another; the keyword sets (*KWS*) approach uses sets of keywords associated with* JSC*s in 1995/1996, and the co-occurrence of keywords from different sets after 1996; and the affiliation patterns (*AFP*) approach uses textual patterns which match affiliations associated with fields of research, and the co-occurrence of patterns in papers. The sets of publications which resulted these three search strategies were assessed using additional data, such as journal names, titles and cited publications. After a final validation by experts, a total of nine converging research areas have been detected using the* KWS *and* FFR *approaches, and none with the* AFP *approach.*

7.1 Introduction

In one of its February 2008 issues, the Lancet published an article on the outcomes of a trial which studied the effects of a treatment with probiotics on patients suffering from severe acute pancreatitis (Besselink et al., 2008). The most significant outcome of this study was an increased risk for mortality in the group that was supplied the

---

* Originally published in Research Evaluation 19(1), 19-27, 2010.
Reproduced with permission of the publisher.

probiotics, and the study was stopped prematurely. In the same issue, the Lancet published an editorial bearing the suggestive title “Probiotics or con”, which was devoted to this grave outcome and its larger consequences. The editorial concluded that without further research, it was no longer tenable to regard the use of probiotics in consumer products as completely risk-free. This trial and its editorial response illustrate an interesting interdisciplinary phenomenon. First, probiotics, which are well-known food additives, were included in the treatment of severely ill people. Next, following the suggestion of the Lancet editorial, probiotics as food additives should now be scrutinized due to the outcome of this application. In a sense, this would complete an interdisciplinary circle, where a result (probiotics) from one field (Food Science and Technology) influences research (trial study) in another field (Clinical Medicine) and vice versa. Developments such as this, where distinct research areas start to apply problems and tools from one another, possibly leading to new research directions or even new research areas, are considered examples of converging research in this paper. More formally, we define converging research as the emergence of new interdisciplinary research from fields which showed limited mutual interdisciplinary connections before. In this definition, emergence is understood as the appearance of thematically related research which shows a significant growth in the number of publications. Also, the converging research is the result of research in two areas which had limited interdisciplinary connections before, and is thus both new interdisciplinary research, and a novel combination of disciplines. Our definition explicitly excludes intra-disciplinary emergence (i.e. within fields), and emergence of this type was not considered in our research.

The publication of the report on the NSF/DOC-sponsored workshop “Converging Technologies for Improving Human Performance” (Roco and Bainbridge, 2003), was important for the widespread introduction of the notion of converging research. There, it was presented as a conceptual framework based on a “unification of science in nature” which allows the creation of mutually enabling research areas with “synergistic effects” (*ibid*, p.1). The workshop focused on the four fields Nanotechnology, Biotechnology, Information Sciences and Cognitive Sciences (the NBIC fields), and the application to the improvement of human performance in general. The European response to this workshop was the assembly of an expert group. This group published a report (Nordmann, 2004) titled “Converging Technologies for a Knowledge Society” (CTEKS). This report stated that converging research was not limited to the NBIC fields, but could originate from anywhere in science, and that the application was generalised to current societal, economical and scientific problems (instead of the improvement of human performance). However, both the NBIC and the CTEKS reports stressed the importance of the research in the Social Sciences and Humanities to investigate aspects such as social and ethical consequences of the application of the research.

The possibility to apply research developments on current problems shows the potential use converging research can have for policy makers. However, just what sets converging research apart from other interdisciplinary developments is not always clear. Expanding our knowledge about this issue was one of the objectives of an

EU project titled “New Emerging/Converging Clusters of Science and Technology”. The project’s most important goal was to identify and describe converging research addressing current societal, economical or scientific problems using publication data. Even more, these data should not only focus on well-covered fields such as the Life Sciences or the Material Sciences, but also include those from Social Sciences and Humanities. In this paper we will review the methodology developed during the EMCOTEC project and reflect upon the results as well as on some of the design choices made.

Converging research is defined as a case of emerging research of a novel, interdisciplinary nature. Many examples exist of descriptions and analyses of emerging research areas with both disciplinary and interdisciplinary roots, like for example Mathematical Logic (Berg and Wagner-Döbler, 1996), Superstring Theory (Budd and Hurt, 1991), elements of Spectroscopy (Ródenas-Torralba et al., 2006), Bioelectronics (Hinze, 1994) or Nanoscience and Nanotechnology (Schummer, 2004). Also, general methodologies to find emerging patterns exists, like those described by Morris (2005), Shibata et al. (2008) and Takeda and Kajikawa (2009). However, like the descriptions of the emerging research areas, these general approaches are applied to a limited set of publications or with a specific topic in mind. Moreover, none of these publications deal with the general and broad search for converging research as defined above.

7.2 Data and Method

This paper presents an approach to find converging research. The development of this approach faced some important challenges:

- The many possibilities to combine research from different parts of science;
- The many ways in which one area can influence another, such as cognitively, socially or methodologically;
- The subjective aspects of core notions such as “research area”, or “interdisciplinary connections”;
- The difficulty in generalising known examples of converging research, so that finding one does not necessarily help finding others;
- The unknown number of examples of converging research, making it very difficult to quantify search performance in terms lqike recall and precision.

The resulting approach is a combination of both quantitative and qualitative aspects, which we will explain in more detail below.

The quantitative aspects use publication and citation data extracted from the Web of Science (WoS) database, which is available to the CWTS under license from Thomson Reuters. The publication data consisted of elements such as source (journal) names, titles and abstracts, affiliations, and keywords for publication published between 1990 and 2005. The citation data consisted of all citations from these publications to publications also covered by the WoS.[61]

We identified converging research in a three-step process: search, inspection and validation by experts. Search was performed by using three strategies described below, which differed in the representation of research field and interdisciplinary connections: the field-to-field references approach; the keywords-sets approach; and the affiliation patterns approach. All strategies tested the development of interdisciplinary connections between pairs of fields. The selected pairs were inspected by using additional publication data to understand what the potentially converging research between these pairs was about. Finally, pairs which appeared to represent converging research, were subsequently shown to experts and only accepted if they also recognised it as such. Before we explain the individual approaches in more detail, together with basic notions about growth and size. As will become clear, the strategies show variation in the implementation of these basic notions, because details of the approaches were developed and fine-tuned during the project, and experience gained from the application of one strategy was used in the fine-tuning of the next.

The above definition makes convergence a case of (interdisciplinary) emergence. Therefore, a notion of significant growth is needed. In our project, we focused on fast, non-linear growth, since this type has in many studies been associated with emerging research themes (Tabah, 1999; Goffman, 1971 and O'May, 1966). One of the tools we developed to identify such growth, which we called the Relative Total Growth (RTG), is the absolute difference between the number of papers in the first and last year, divided by the total number of papers. Since smaller values of the RTG are either due to a small difference between the first and last years, or to a small change relative to the total number of papers, important examples of fast growth can be found at the top of a list sorted on descending RTG. However, this indicator is not perfect, and for example does not take into account the shape of the development between the first and the last year; therefore, the RTG was only used to sort promising results.

A research area should show a basic, noticeable level of activity, which should also be measurable in number of publications. However, it is very difficult to quantify "a basic level" exactly. Looking at examples of emerging research areas described in the literature, a total number of publications published per year at the end of the development, can be for example around 100 for Fuzzy Set theory (Berg and Wagner-Döbler, 1996), or around 250 for Bioelectronics (Hinze, 1994). Moreover, the number of publications involved in earlier years can be small: even less than 10 publications per year (*ibid*). The search strategies described below used different lower bounds on the number of publications, due to the experience gained in applying earlier strategies.

We now describe the three strategies in more detail. The most important difference between these strategies is the way they characterise research fields (sets of papers with related research), as well as interdisciplinary connections (use of research from different areas). In the different characterisations, we have tried to span different

---

[61] The WoS database available to the CWTS contains publication data from 1980 and onwards.

dimensions of the science system: references represent (among others) the social and cognitive dimensions, the social and geographical dimensions, and keywords the cognitive and vocabulary dimensions.

Field-to-field references (FFR)

The first strategy we developed is the field-to-field references strategy (FFR). It characterises research fields as Journal Subject Categories (JSCs), which assign each journal in the WoS to at least one and at most six categories. These JSCs are part of the WoS and at the time our research was conducted, there were 243 JSCs. Although this classification is not perfect, it provides a clear, fixed and consistent field definition suitable for automated procedures (Van Raan, 2008). Nevertheless, problems exist, especially with the differences in size and (breadth of) scope, and the multiple assignments of journals to different JSCs. However, such multiple assignments can also be regarded as hubs between categories, which was used by Morillo et al. (2003) to develop an indicator for interdisciplinarity. The FFR strategy characterised interdisciplinary connections by the references from one category to another. To give an interesting example of this, the JSCs Religion and Allergy are connected through a citation in 2000 from the journal Zeitschrift für Evangelische Ethik to the journal International Archives of Allergy and Immunology by Haniel (2000) citing Hammer et al. (1998). Our strategy excludes author self-citations and journal self-citations, since these may be indicative for other processes than use or exchange of knowledge (Price, 1981a). Our analysis covered 15 years of publication data (1991-2005). The lower bound on the total number of references was set to at least 200 references, employing a three-year citation window.[62] Additionally, the growth should be positive for the last five years, i.e., every subsequent year should have more references than the year before. Finally, the last two years should show a growth that is either larger than the average growth in the previous year, or larger than the average growth the in the last two years of the all other remaining pairs.

Affiliation patterns (AFP)

The second search strategy characterises research fields by textual patterns for the affiliations (addresses) associated with publications. The list of patterns was created as follows. For each JSC, frequency lists of affiliation names were collected. Based on these lists, patterns like “econ*” for Economics or “fam*” for Family Studies were devised manually. Additionally, related JSCs were combined and used one affiliation pattern, resulting in a 121 patterns for 243 JSCs. Interdisciplinary connections are characterised as the co-occurrence of different patterns in the list of affiliations associated with a single paper, like in Gundersen and Ziliak (2004) where Department of Human Development and Family Studies of the Iowa State University co-authored

---

[62] A citation window is the maximum number of years between the citing and the cited publication.

with the Department of Economics of the University of Kentucky. Experience with the FFR strategy caused a reconsideration of the lower bound on the number of publications, notion of significant growth and the number of years in the analysis: a single year with at least ten publications was used as the lower bound, the RTG was used to rank the most promising ones at the top, and we now use eleven years of data (1995–2005), because we focus on recent developments. The choice for ten publications was based on the inspection of the distribution of the RTG, which appeared to resemble a normal distribution[63] without too many outliers when at least ten publications were required. The inspection of the sorted list was stopped after inspecting 50 pairs.

Keyword sets (KWS)

The third and final search strategy characterises research fields as sets of keywords. The keywords we use are the "keywords plus", which are associated with many of the publications in the WOS (Garfield and Sher, 1993). For each JSC, keywords are selected based on publications published in the first two years (1995–1996). The objective of this selection is to catch the cognitive core of the fields, i.e. the content which differentiates the fields. To accomplish this, the top 20% most frequently used keywords is selected first. Then, to reduce overlap between sets, only keywords were kept which had at least 50% of their covered publications in one JSC. Finally, to reduce the processing requirements, we selected at most 500 of the keywords sorted on ascending coverage percentage and descending frequency. Based on our experience, 500 can be considered large enough. Additionally, we did not create sets for multidisciplinary JSCs as well as general, interdisciplinary JSCs, because early inspection revealed that such categories had too many words in common with too many other categories, and negatively influenced the set creation.

Inspection

We inspected selected pairs as follows. First, we dismissed pairs judged to be related a priori, usually pairs of JSCs with similar scope. Such judgements used objective criteria, like the number of journals in common, but also background knowledge. Background knowledge also was also used to dismiss "old convergence", i.e. pairs we knew to represent the result of convergence, but which took place some time ago. Next, the content of remaining pairs was inspected, with the objective to find a limited number of subjects that would indicate the research focus of the convergence. Candidates would be dismissed if too many subjects were found or if the subjects were too unrelated. Finally, the candidates were given tentative names that reflected their research focus and passed on to experts.

---

[63] This was based on visual inspection and not verified statistically.

7.3 Results

7.3.1 Application of the search strategies

In this section we present the application our search strategies and the inspection and validation of the results. The combined result of all strategies is given in **Table 5**.

Research Area	**Strategy**	**#**	**Category 1**	**Category 2**	**N**
			Accepted		
Biomaterials	FFR	1	Chemistry, Multidisciplinary	Engineering, Biomedical	8,806
		2	Chemistry, Multidisciplinary	Materials Science, Biomaterials	
		3	Chemistry, Physical	Materials Science, Biomaterials	
Biosensors and Biodevices	FFR	1	Biochemical Research Methods	Electrochemistry	2,982
ICT Networks	FFR	1	Computer Science, Theory & Methods	Telecommunications	8,474
Health Aspects of Nutritional Components	KWS	1	Food science & Technology	Endocrinology & Metabolism	224
		2	Nutrition & Dietetics	Gerontology	
Mild Cognitive Impairment and Cerebrovascular Disease	KWS	1	Peripheral Vascular Disease	Neuroscience	60
Neuro Imaging and Brain Imaging	KWS	1	Radiology, Nuclear Medicine & Medical Imaging	Psychology, Experimental	215
Polymers and the Central Nervous System	KWS	1	Polymer Science	Neuroscience	75
Quantum Information Science	KWS	1	Polymer Science	Engineering, Electric & Electronic	145
Surface Physics and Chemistry for Biological and Medical Applications	KWS	1	Physics, Applied	Dentistry, Oral Surgery & Medicine	107
			Too old		

Table 5.a The research areas identified using the different strategies, together with their expert assessment and the candidate pairs grouped together into one area. The overlap was calculated as the total number of (citing) publications published between 1995 and 2005.

Research Area	Strategy	#	Category 1	Category 2	N
Bioinformatics	FFR	1	Biochemistry & Molecular Biology	Statistics & Probability	22,999
		2	Biotechnology & Applied Microbiology	Statistics & Probability	
		3	Computer Science, Interdisciplinary Applications	Genetics & Heredity	
		4	Computer Science, Theory & Methods	Biotechnology & Applied Microbiology	
		5	Genetics & Heredity	Mathematics, Interdisciplinary Applications	
		6	Statistics & Probability	Biotechnology & Applied Microbiology	
Cancer Supportive Care	KWS	1	Clinical Psychology	Oncology	67
Psychological Distress in People with Diseases of the Respiratory System	KWS	1	Respiratory Systems	Psychiatry	54
			Dismissed		
Genetic Deficiencies and Dermatological Diseases in Relation with Neurological Phenomena	KWS	1	Dermatology	Clinical Neurology	54
Glasses for Fibre Lasers and Amplifiers	KWS	1	Materials Science, Ceramics	Engineering, Electrical & Electronic	25
Simulation and Modelling of Nanotubes and Hollow Fibre Membranes	KWS	1	Microbiololgy	Agriculture, Soil Science	55
		2	Chemistry Analytical	Agriculture, Soil Science	
Socio-Psychological Aspects of Socio-Economic Changes and Circumstances	KWS	1	Developmental Psychology	Economics	91
Soil Microbiology and Soil Analytical Chemistry	KWS	1	Mathematics	Engineering, Chemical	86

Table 5.b The research areas identified using the different strategies, together with their expert assessment and the candidate pairs grouped together into one area. The overlap was calculated as the total number of (citing) publications published between 1995 and 2005.

Field-to-Field References

About 37,000 pairs contained at least one reference from the citing to the cited category, and about 5,900 pairs had at least 200 references in 15 years. Applying our growth requirements reduced this to 38 candidate pairs. Of these 38 pairs, those that had too many journals in common were dismissed, because we considered such pairs as representing well-known converged areas (for instance Nanotechnology). The final selection consisted of 11 pairs. Some of these pairs showed enough overlap to combine them in a single research area. Consequently, the 11 pairs were grouped into four converging research areas. In **Table 5** these four areas are indicated with the FFR label in the strategy column.

Affiliation Patterns

About 5,200 pairs of patterns had at least one publication in common and almost 3,000 pairs had at least one year with ten or more publications. Unfortunately, no research focus could be found in the top 50 pairs sorted on descending RTG. Instead, publications referred only research in one of the fields, or to research in fields that appeared not related to the affiliation, or contained a range of subjects instead of only limited number. For example, the pair Opthalmology (opth*) and Neuroimaging (neuroimag*) was a selected candidate, but the associated publications dealt with variety of (in our opinion) unrelated topics, such as “fatal insomnia” and “kinase inhibitors”.

Keyword Sets

Over 7,500 pairs of sets had at least one publication in common and almost 3,000 pairs had at least one year with ten or more publications. These pairs were sorted on descending RTG and the top 50 was inspected. After inspection, we selected 15 pairs. Again, some of these pairs again showed overlap and were grouped together. The final result consisted of 13 research areas. In **Table 5** these areas are indicated with a KWS in the strategy column.

Expert verification

The combined result from the FFR and KWS strategies consisted of seventeen research areas, together with a description of their content. This result was presented to field experts, who were asked if they recognised the research area and if so, whether the area was the result of a convergence of other areas. Additionally, they were asked to verify or improve the (tentative) names given to the areas. The verification resulted in nine confirmed research areas, three areas judged too old, and five dismissed research areas, either because they were not recognised or were not the result of convergence. This result is reflected by the different parts “Accepted”, “Too old”, and “Rejected” in **Table 5**.

7.3.2 Additional Analyses

The differences in the overlap column of **Table 5**, which is representative for the size of the interdisciplinary area, are immediately apparent. In the case of the FFR

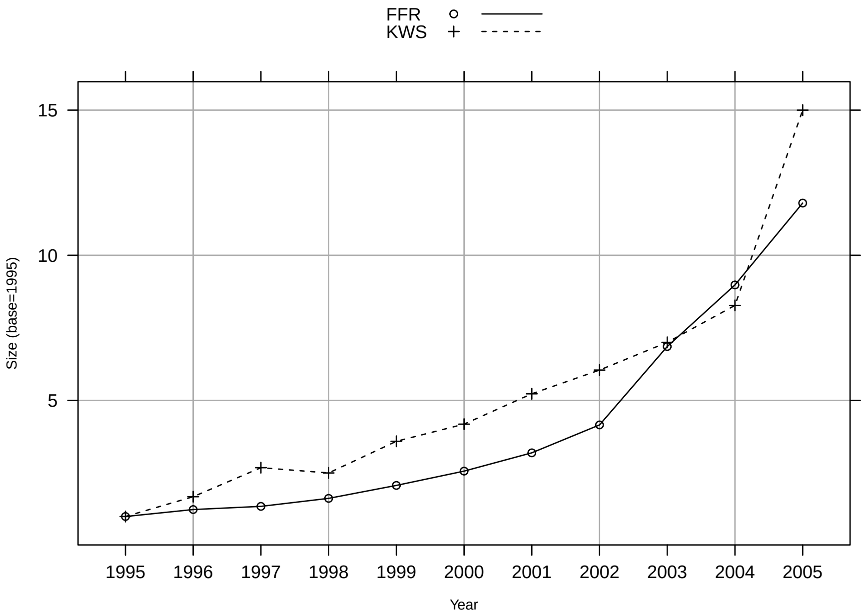


Figure 19 The combined number of publications involved in the overlap between research areas (**Table 7**) indexed to 1995.

strategy this was measured by the number of citing publications, while in case of the KWS strategy it was measured by the number of publications associated with both keyword sets. One can also note this difference in size by the names given to the research areas, which are shorter and more general for the larger areas, identified using the FFR strategy, while they are longer and more specific for the smaller KWS areas. To further investigate this difference, **Table 6** contains a summary of the distribution of the number of publications in the overlap, measured between 1995 and 2005. This table also includes the numbers for the AFP strategy, because these numbers could hint at reasons for the failure of the AFP approach. The table shows that the coarseness of the categories and the affiliation patterns are comparable, but that the spread of the size is wider in case of the affiliation patterns. Additionally, it shows that the size of a publication set covered by a keyword set is on average about 20% of that of a category, and about 18% of that of an affiliation pattern. However, the difference measured by the median is even larger, and then the size of the keyword sets is only respectively 9% and 13%. Furthermore, for the FFR and KWS approach, **Table 7** shows the size from year to year. These numbers are also shown in **Figure 19**, but normalised by setting the value in 1995 to 1.0. The figure shows that, although the difference in absolute numbers of papers in **Table 7** is large, the development of the growth has a comparable shape for both strategies.

Characterisation	**Min**	**Median**	**Mean**	**Max**
FFR	3,243	48,370	78,180	654,700
AFP	20	31,450	88,160	1,115,00
KWS	13	4,368	16,080	154,70

Table 6 A summary of the size of the different research area characterisations, measured by the number of (citing) publications published between 1995 and 2005.

Strategy	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
FFR	965	1,196	1,302	1,568	1,999	2,474	3,082	4,010	6,616	8,664	11,385
KWS	22	37	59	55	79	92	115	133	154	182	330

Table 7 The combined number of publications involved in the overlap between research areas.

A final additional analysis is given in **Table 8**, which shows the major science fields involved in the validated converging research areas. This table was created by taking the JSCs involved in the research area characterisations, and aggregating those to their major science fields.[64]. From **Table 8** it is clear that the Life Sciences and the Natural Sciences dominate the results, while only one out of nine areas involved work from the Social Sciences and two out of nine involved work from Engineering. Even more, the two latter fields are never without JSCs that are also associated with the Life Sciences or the Natural Sciences.

Research area	**Strategy**	**Main fields**			
		Life	**Natural**	**Engineering**	**Social**
Biomaterials	FFR	Yes	Yes	No	No
Biosensors and Biodevices	FFR	Yes	Yes	No	No
ICT Networks	FFR	No	Yes	Yes	No
Health Aspects of Nutritional Components	KWS	Yes	No	No	No
Mild Cognitive Impairments and Cardiovascular Diseases	KWS	Yes	No	No	No
Neuroimaging and Brain Imaging	KWS	Yes	No	No	Yes
Polymers and the Central Nervous System	KWS	Yes	Yes	No	No
Quantum Information Science	KWS	No	Yes	Yes	No
Surface Physics and Chemistry for Biological and Medical Applications	KWS	Yes	Yes	No	No
Total involvement		7	6	2	1

Table 8 Main fields of science involved in the detection of the selected research areas.

---

[64] A CWTS taxonomy is being maintained which assigns JSCs to science fields and major fields

7.4 Discussion

We presented three strategies for locating converging research using publication and citation data from the WoS database. Converging research was defined as the emergence of an interdisciplinary area from areas between which limited interdisciplinary connections existed before. This type of emerging research has not been studied extensively using publication data before, and it was not known beforehand what the different strategies could yield. It is therefore encouraging to see that our work yielded 26 converging pairs, which were grouped into seventeen research areas. Nine of these passed the judgement of experts, while three others were confirmed but dismissed as "too old". It was found that a large size difference exists between the converged pairs resulting from the FFR strategy and the KWS strategy. However, the shape of the developments in the resulting areas was found to be comparable for both strategies. Finally, the results were mainly from the Life Sciences and the Natural Sciences.

It is difficult to judge which strategy was more fruitful: although more areas were identified using the KWS strategy (thirteen in total), more than half (eight) of these areas were considered too old or were otherwise dismissed. On the other hand, only four areas were identified using the FFR strategy, yet only one was dismissed as too old. At the same time it is not clear why the AFP strategy could not help to identify converging research: although patterns of fast growing co-occurrences of different affiliation patterns were found, none appeared to focus on the converging pairs, or on a limited number of subjects. Together with the distribution of the size of the overlap between affiliation patterns in **Table 6**, this may suggest that the current patterns are too multidisciplinary and may fail to provide enough a priori focus. To give an example, the JSC Crystallography has a maximum overlap of 8% with other JSCs, but the affiliation pattern for Crystallography fails to reflect this: only about 25% of the papers are associated with only the JSC Crystallography and the rest with other JSCs. One could argue that this may indicate a multidisciplinary nature of the research in Crystallography, but it also appears not to provide a useful starting point to search for converging research. Perhaps this is for the same reasons for which we excluded the multidisciplinary JSCs in the KWS strategy, but it nevertheless deserves more attention in future research.

7.4.1 Review of the methodology and future research

An important advantage of our approach is that it has a quantitative basis. Therefore, it becomes possible to repeat and update the search with new data or fine-tuned parameters. Additionally, the application of a search strategy on the whole of science avoids a thematic bias and therefore improves the objectiveness of results. Other approaches, such as interviewing experts, may run the risk of a possibly biased view. On the other hand, experts may be able to signal developments sooner, since they may be part of specific developments, and are therefore not limited by the publication delays that are inherent to approaches based on peer-reviewed publication data.

However, it is not an empirical fact that peers are always more up to date than our approach.

Some aspects in the strategies can be improved. Already during the project, the short-term focus provided by the three-year citation window used in the FFR approach was considered too short, given that interdisciplinary knowledge transfer takes time (Rinia et al., 2001a, 2002). Also, a lower limit of 200 citations was high and possibly resulted in a bias toward large (biotechnological) fields. Improved versions of the strategies should select pairs on the basis of growth characteristics instead of (only) size characteristics. Also, the three strategies are (directly or indirectly) based upon the JSCs. Although the JSCs have certainly been extremely convenient in the development of the strategies, alternative groupings could also be employed. For example, the use of smaller groups of journals based on journal citation patterns could improve both the FFR and the KWS strategies. The KWS strategy "freezes" the keywords in the years 1995 and 1996. This was a conscious design choice, since this limited the use of the JSC on which it was based to only those years. However, this implies that knowledge developed after 1995 and 1996, which may use new and not included keywords, is disregarded and newer developments are not included. Another problem of the KWS strategy is due to homonymy and synonymy. For example, it is not possible to distinguish between AD as an abbreviation for Alzheimer's disease or for Advertisement. Finally, the removal of the multidisciplinary categories was another design choice, because those categories made the distinction between other fields less clear, clouding the interdisciplinary developments. However, multidisciplinary JSCs can also be regarded as a bridge between fields and a basis for new interdisciplinary developments. Therefore, future research should focus on how to reintegrate them, or to use them as a source of information.

The current approach requires a qualitative selection process based on a judgement of resulting publications. Although this introduces a subjective element in the search, this also allows the process to include background knowledge which is difficult to express in quantitative terms alone. The use of background knowledge is witnessed by the small number of publications in some of the KWS pairs, which were chosen because both we and the experts recognised them to represent known research, without covering all publications associated with that research. At the same time, we also dismissed pairs as too old using this kind of background knowledge. Additional tools could be used to try to encode such background knowledge and use that in the search and inspection. For example, Buter et al. (2006) describes an interface between a (quantitative) cognitive science map, which encodes the content of a set of publications, and a qualitative, user-provided concept map that encodes background knowledge in the form of research topics.

Another extension which can be used in the inspection phase of the approach is to expand the number of publications by those which share an important part of the knowledge base (such as the best cited publications). Such an extended set could create a more complete picture of the focus of the research in the convergence. To establish coherence of the (expanded) publication sets and aid the search for focus in the research, visualisation tools can help, such as cognitive maps (Buter and Noyons, 2001, 2002a) or networks (Calero et al., 2006 and Takeda and Kajikawa, 2009).

7.4.2 On the nature of converging research

We close this discussion with thoughts on the nature of converging research, combining theory and results described above. The original idea of converging research Roco and Bainbridge (2003) was more of a conceptual, high-level nature, rather than a concrete phenomenon. It was based on the idea that a unification of the sciences based on a "unity in nature" (*ibid*, p.15) could enable new developments. This conceptual framework was put in more concrete terms in this paper by defining converging research as emerging interdisciplinary research areas. Also, there were no explicit or implicit restrictions put on the fields that could converge; and in principle any (exotic) combination was possible. However, many of the pairs found in **Table 5** appear already related, even though they are different fields without shared journals or (initially without) shared keywords. For example, it appear not surprising that Multidisciplinary Chemistry may use results from Materials Science, Biomaterials; or that Biochemical Research Methods cites Electrochemistry. This also confirms earlier observations reported by Porter and Chubin (1985) who also found that the use of research results from distant fields is a rarity, using a limited set of journals and categories.

A possible reason for this is that such endeavours have an implicit risk for the researcher. So can importing knowledge from a completely different domain may be less productive (Palmer, 1999). Additionally, if a researcher ventures into a new community, there is a need to learn and conform to existing vocabulary, descriptions of tools as well as problems. Acceptance and recognition by a new community may also be a boundary for a successful cross-over. Successful convergence therefore requires a mode of working that has a common language or understanding as well as a way to provide recognition of researchers from both originating disciplines. A "unity in nature" is usually not enough. At the same time, (*ibid*) also notes that importing knowledge from other domains offers a better opportunity for knowledge base development, and provides researchers with the opportunity to distinguish themselves by applying methodology to a new problem domain or (vice versa) to attack a problem with new methodology. This is also what we see in our results: science-born applications of tools on problems, where both tools and problems can be imported. For instance, the JSC Computer Science, Theory and Methods imports problems from the JSC Telecommunications (related to networks), and applies its own methods in the converging research area. In the Bioinformatics converging research area however, the Genetics and Biochemistry fields import tools from Statistics, for instance in order to apply these tools to expression data from micro-arrays.

Such developments result from the creative "probing" efforts of researchers into new research directions, or caused by interdisciplinary developments related to tasks, processes, products, or use (Porter and Chubin, 1985). Those processes are part of normal science and are not necessarily indicative for Kuhnian paradigm shifts.

7.5 Concluding Remarks

The results presented in this paper show that converging research can successfully be identified using search strategies applied to publication and citation data. Although convergence was originally developed on a more conceptual level, the translation into detectable, scientific developments makes it an interesting evaluation tool, because it can highlight new interdisciplinary research developments with a novel character. These developments are interesting for at least two reasons: first, converging research areas are spontaneous, normal developments in science; and second, convergence occurs between fields that have some kind of conceptual distance between them. Especially the last point makes convergence a potentially valuable phenomenon, because combining knowledge domains requires creativity and effort, and is also not without risk. Therefore, to invest such effort and run such risks, potential benefits must appear valuable. At the same time, this also implies that most convergence was (and will be) found between fields that are not too distant from each other, since the effort to cross fields increases proportional to the conceptual distance between them.

We conclude that, even though our strategies still require more development and the process as a whole would benefit from additional tools, the quantitative core of our approach could serve as the basis for a monitor that would signal new converging developments. Such a monitor could be a part of a policy instrument to identify areas where tools and problems are combined in new and creative ways, some of which may help to address current social, economical and scientific issues.

Acknowledgements

The authors of this paper are grateful for the work done by our colleagues in the EMCOTEC project: Christien Enzing and Marc van Lieshout from the Netherlands Organisation for Applied Scientific Research (TNO), and Vera Grimm and Dirk Holtmannspötter from the Association of German Engineers (VDI). Also, we are indebted to the participants of the final EMCOTEC conference in Brussels (March 2008) who helped us shape our thoughts with their presentations and discussions. Finally, this work was done with support from the EU, as it was part of the "New Emerging/Converging Clusters of Science and Technology" (EMCOTEC) project, contract SDPF-CT-2005-00002.

References

Berg, J. and Wagner-Döbler, R. (1996). A multidimensional analysis of scientific dynamics. *Scientometrics, 35*(3), 321–346.

Besselink, M. G. H., Van Santvoort, H. C., Buskens, E., Boermeester, M. A., Van Goor, H., Timmerman, H. M., Nieuwenhuijs, V. B., Bollen, T. L., Van Ramshorst, B., Witteman, B. J. M., Rosman, C., Ploeg, R. J., Brink, M. A., Schaapherder, A. F. M., Dejong, C. H. C., Wahab, P. J., Van Laarhoven, C. J. H. M., Van Der Harst, E., Van Eijck, C. H. J., Cuesta, M. A., Akkermans, L. M. A. and Gooszen, H. G. (2008). Probiotic prophylaxis in predicted severe acute pancreatitis: a randomised, double-blind, placebo-controlled trial. *The Lancet, 371*(9613), 651–659.

Budd, J. and Hurt, C. D. (1991). Superstring theory: Information transfer in an emerging field. *Scientometrics, 21*(1), 87–98.

Buter, R. K. and Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68.

Buter, R. K. and Noyons, E. C. M. (2002a). Using bibliometric maps to visualise term distribution in scientific papers. In *Information Visualisation, International Conference on*, pages 697. Los Alamitos, CA, USA.

Buter, R. K., Noyons, E. C. M., Van Mackelenbergh, M. and Laine, T. (2006). Combining concept maps and bibliometric maps: First explorations. *Scientometrics, 66*(2), 377–387.

Calero, C., Buter, R. K., Valdés, C. C. and Noyons, E. C. M. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics, 66*(2), 365–276.

Garfield, E. and Sher, I. H. (1993). Keywords plus algorithmic derivative indexing. *Journal of the American Institute for Information Science, 44*(5), 298–299.

Goffman, W. (1971). A mathematical method for analyzing the growth of a scienfic discipline. *Journal of the Association for Computing Machinery, 18*(2), 1672–1673.

Gundersen, C. and Ziliak, J. P. (2004). Poverty and macroeconomic performance across space, race, and family structure. *Demography, 41*(1), 61–86.

Hammer, Linke, Wagner and Diefenbeck (1998). Organs from animals for man. *International Archives of Allergy & Immunology, 116*(1), 5–21.

Haniel, A. (2000). Organs at any price? on the question of alternatives to post-mortem organ donation (examining ethical aspects of human organ cultivation and cloning). *Zeitschrift für Evangelische Ethik, 44*(4), 269–284.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: the case of bioelectronics. *Scientometrics, 29*(3), 353–376.

Morillo, F., Bordons, M. and Gómez, I. (2003). Interdisciplinarity in science: a tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology, 54*(13), 1237–1249.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, cocitation and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, 56*(12), 1250–1273.

Nordmann, A. (2004). Converging technologies — shaping the future of European societies. Technical Report EU HLEG Foresighting the New Technology Wave.

O'May, K. (1966). Quantitative growth of the mathematical literature. *Science, 154*, 1672–1673.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science, 50*(3), 242–253.

Porter, A. and Chubin, D. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3), 161–176.

Price, D. D. S. (1981a). The analysis of scientometric matrices for policy implications. *Scientometrics, 3*(1), 47–53.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2001a). Citation delay in interdisciplinary knowledge exchange. *Scientometrics, 51*(1), 293–309.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics, 54*(3), 347–362.

Roco, M. C. and Bainbridge, W. S., editors (2003). *Converging Technologies for Improving Human Performance*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ródenas-Torralba, E., Morales-Rubio, Á. and Guardia, M. D. L. (2006). Scientometric picture of the evolution of the literature of automation in spectroscopy and its current state. *Spectroscopy Letters, 39*, 513–432.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics, 59*(3), 425–465.

Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*(11), 758–775.

Tabah, A. N. (1999). Literature dynamics: studies on growth, diffusion and epidemics. In Williams, M. E., editor, *Annual Review of Information Science and Technology*, pages 249–286. Information Today Inc.

Takeda, Y. and Kajikawa, Y. (2009). Optics: a bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics, 78*(3), 543–558.

Van Raan, A. F. J. (2008). Bibliometric statistical properties of the 100 largest european research universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology, 59*(3), 461–475.

8

Searching for converging research using field to field citations*

Abstract

We define converging research as the emergence of an interdisciplinary research area from fields that did not show interdisciplinary connections before. This paper presents a process to search for converging research using journal subject categories as a proxy for fields and citations to measure interdisciplinary connections, as well as an application of this search. The search consists of two phases: a quantitative phase in which pairs of citing and cited fields are located that show a significant change in number of citations, followed by a qualitative phase in which thematic focus is sought in publications associated with located pairs. Applying this search on publications from the Web of Science published between 1995 and 2005, 38 candidate converging pairs were located, 27 of which showed thematic focus, and 20 also showed a similar focus in the other, reciprocal pair.

8.1 Introduction

Some say that ‘a Jack of all Trades is a master of none’. Yet, the days in which “[...] our universities [...] [were] divided into different departments that [did] not know very much of each other” (Cassirer, 1942, p.309), appear behind us. Instead, interdisciplinary research “is spreading all over the landscape of science and technology” (Gibbons et al., 1994, p.22) for more than one reason. For instance, knowledge migration is a fruitful mechanism by which science expands into new realms (De Mey, 1982, p.140-145), providing an attractive opportunity for researchers to attain recognition. Also, interdisciplinary research has been encouraged by funding agencies as a problem-driven mode of research (Carayol and Thi, 2005), or as a way to “attain

---

* Originally published in Scientometrics 86(2), 325-338, 2010.
Reproduced with permission of the publisher.

[...] system solutions" for complex societal problems (Persson, 1999). In spite of this encouragement, the status and prestige of interdisciplinary research is not clear-cut. On one hand, Rinia et al. (2001a) found no (general) evidence of a negative bias to interdisciplinary research in peer-review assessments of physics, nor in relative bibliometric indicators. On the other hand however, De Boer et al. (2006) quote researchers who attribute a perceived lack of prestige to interdisciplinary research. Also, Carayol and Thi (2005) found evidence of a lack of incentive in the academic reward system. To add to this uncertainty, Larivière and Gingras (2010) show that the amount of interdisciplinary links in a publication, which is common in many interdisciplinary pub- lications, may influence impact. Apart from the uncertain academic reward of an inter- disciplinary venture, Palmer (1999) notes that even experienced scientists feel challenged by the task of acquiring knowledge from outside their realm of expertise. Additionally, the social, political, and cultural structures of research areas create other barriers (Ruiz-Baños et al., 1999).

In all, we conclude that moving to research outside a familiar setting requires an investment in new knowledge, new vocabulary, as well as new social structures and customs, while the pay-off in terms of acceptance, let alone an increase in status and prestige among peers, appear less certain. This tension makes investigating the creation and subsequent growth and decline of interdisciplinary research areas even more interesting.

We define converging research as the emergence of a new interdisciplinary research area from fields which showed no such interdisciplinary connections before; the result we call converged research, or a convergence. Our definition is similar to that of Nordmann (2004), who speaks of mutually enabling systems and technologies in pursuit of a common goal, but our definition is more targeted to science systems. At the same time, it is more specific than that of Roco and Bainbridge (2003), who speak in more conceptual terms and identify trends on a very large scale: the "megatrends" in Roco (2002). Also, it is similar to the use of convergence as a technological phenomenon (Gambardella and Torrisi, 1998 and Rosenberg, 1976): for example, the shared technical basis underlying convergence in industry can be compared to the generality of research methods or tools.

In this paper, we describe a process that locates converging research based on journal subject categories in the Web of Science database as proxies for fields. Citations (from one field to another) are used to measure interdisciplinary connections. Our working procedure consists of a quantitative and qualitative part: the first part locates candidates, and the second part inspects those candidates. The quantitative part uses an objective basis for the cut-off value for significant (in the sense of "noticeable larger") growth, as suggested in our previous article (Buter et al., 2010a). The qualitative part is based on a visual inspection of a tableau of graphs, as well as an inspection of publication data assembled from the converging area.

Converging research has not been the topic of many scientometric publications, although there are a lot of descriptions and analyses of emerging research, like for example Mathematical Logic (Berg and Wagner-Döbler, 1996), Bioelectronics (Hinze, 1994) or Nanoscience and Nanotechnology (Schummer, 2004). Additionally, investigations of interdisciplinary developments have been described in for example Davidse and Van Raan (1997), Rinia et al. (2002), and Morillo et al. (2003). Also, general methodologies to find emerging patterns have been described by for example Morris (2005), Takeda and Kajikawa (2009), or Upham and Small (2010). However, like the descriptions of the emerging research areas, these general approaches are applied to a limited set of publications, or for a specific topic, and none of these publications deal with the general and broad search for converging research we describe in this paper.

8.2 Data and method

8.2.1 Bibliometric data

The bibliometric data used for our research, consisted of all publications in the Web of Science (WoS) database[65] published between 1995 and 2005, including the social sciences and humanities. We also used the citations from these 1995–2005 publications to WoS publications, but the cited publications were restricted to "articles", "letters", "notes" and "reviews" (the citing documents could be any type). Both author self-citations and journal self-citations were excluded. Author self-citations were excluded because such citations may represent other mechanisms than the use of research (Price, 1981b). Journal self-citations on the other hand, were excluded because our method was based on journal subject categories, and we found such that self-citations introduced noise in detecting interdisciplinary developments. We used all 243 journal subject categories (JSCs) provided by Thomson Scientific in 2005, which categorized all journals in the WoS into at least one and at most six categories.

8.2.2 Significant growth, non-linear shape and robust size

The search method we developed, had the objective to find robust phenomena showing significant, non-linear growth. In this section we describe how we implemented growth, significance, shape and robustness.

Our growth indicator used citation counts within a specific citation window: the range of years in which cited publications are published relative to the publication year of a citing publication. For example, for a publication from 1995 and a citation window of 10 years, we count the citations to publications from 1986 to 1995[66] A small citation window focuses on recent (relative to the citing publication) developments, whereas a longer window starts to include more and more citations to "classics". We

---

[65] his database is available to the CWTS' under license from its publisher Thomson Scientific; it contains publication data from the year 1980 onward.

[66] We count whole years, so 1995–1996 counts as a 2-year citation window, even though the difference is only 1 year.

chose a 10 year wide citation window as a good compromise. Since knowledge transfer takes time (Rinia et al., 2001a), a smaller window may run the risk of not including interdisciplinary usage, while a wider window would include more classics, which we consider of limited use as they are probably too general to mark actual research.

Large differences exist in the average number of citations per publication in different fields: for instance, in our data set, the field Genetics and Heredity had an average of over 33, whereas Mathematics only had about 7. We therefore normalised citation counts from one field to another, by dividing the individual counts by the total number of citations (for a given year) from the citing field to all fields (including itself); we refer to this normalised count as the citation share and denote this as c(A,B)t for year t and fields A and B. For example, if Mathematics would give 100 citations to Genetics and Heredity in 1996, out of a total of 2,000 from Mathematics in 1996, then the citation share of Mathematics to Genetics and Heredity in 1996 would be 0.05. Next, since growth is about change, we used the difference of citation shares in subsequent years $t-1$ and t, divided by the (absolute) value for the previous year. To this we refer to as the growth rate $g(A,B)_t$, and we can capture its definition in a formula as follows:

$$g(A,B)_t = \begin{cases} \frac{c(A,B)_t - c(A,B)_{t-1}}{|c(A,B)_{t-1}|} & \text{if } |c(A,B)_{t-1}| > 0; \\ 0 & \text{otherwise.} \end{cases}$$

This growth rate was used to identify significant, fast-growing growth.

We wanted an objective basis for the distinction between significant and non-significant growth, because a previous version of our search process, described in Buter et al. (2010a), used a strict value which was later considered too large. After some experimentation, significant growth was defined as follows. For all growth-rate time-series of the citing-cited pairs, the median[67] value was calculated. The distribution of these median values appeared to be log-normal and we used this distribution characteristic to define significantly growing time-series to be those which had a median value of at least 1.5 standard deviations larger than the average of this distribution. Similar considerations were used in the definition of the RTG[68] indicator (Buter et al., 2010a), and also in Efron and Tibshirani (1993) to assess the significance of the differences between two groups of values.

To select fast-growing pairs in those with significant growth, we first experimented with methods which fit non-linear curves, as well as smoothing functions such as described by Silverman (1985). Unfortunately, the results were not useful, most probably because the time-series of growth rates were small (only ten observations for the years 1995–2005), as well as coarse (large variation of values in a time-series).

---

[67] We used a median instead of a mean value, because the median is less affected by the outliers present in many of our time-series of growth rates due to the limited number of values in the series.

[68] This number is defined as the absolute difference between the number of papers in the first and last year, divided by the total number of papers; it has an approximately normal distribution if the required number of publications is large enough.

For the same reasons, indicators such as developed in Egghe and Rao (1992) did not yield useful results. Consequently, we resorted to a more basic approach, and devised two straightforward requirements which expressed our interested in recent, fast growth. First, the maximum value of the growth rate should fall between 2002 and 2005. Second, the sum of the citation counts in the period 2001–2005 should be double the sum of counts in the previous period (1995–2001).

A subject can be called to show a robust development if it has a "large enough" number of publications in order to be interesting. However, a requirement such as "large enough" is quite subjective and difficult to express exactly. As a result, there is a level of arbi- trariness in the two robustness requirements we used, but we consider them quire acceptable: first, more than half (six) of the values in the growth rate had to be larger than 0; and second, at least 1 year in the period 2002–2005 had to have 25 citations or more. Again, some experimentation was required in order to find these values.

8.2.3 Assessment

The objective of the assessment was to find out more about citing and cited publications of the pairs found after applying the above requirements. Although our main concern was to find common (research) themes, we were also interested in a graphical display of citation counts, in order to evaluate our search method by verifying that our requirements did indeed result in the desired growth shapes.

The shapes of the citation counts and shares were inspected using a tableau of graphs similar to the one shown in **Figure 20** for the pair Economics citing Physics, Fluids and Plasmas. This tableau is divided vertically in two: the left contains graphs for a pair selected by the search (Economics citing Physics, Fluids and Plasmas), and the right part contains the same graphs for the reciprocal pair, in which the citing and cited field are exchanged (Physics, Fluids and Plasmas citing Economics). The graphs at the top show the citation share time-series. Below those, the time-series of absolute citation counts are plotted. In order to see any effects of citation delay, also plots for citation windows of different sizes (3, 5 and 10 years) are plotted. Finally, in order to rule out results that are due to a sudden increase in citations in a field as a whole[69], graphs are plotted at the bottom which compare the (scaled) citation counts of a pair with the (scaled) number of citations given in both originating fields.

In this paper, research focus is understood as the most specific common theme present in most of the publications under inspection. This is a liberal definition and it includes common themes in research subjects as well as in applied methodology. However, we do require research focus to be the most specific common theme, as we expect to find multiple themes in many of the pairs we find. The research focus of the pair located by our search method was leading, and alternative focus in the reciprocal

---

[69] For example, due to the inclusion of a new journal with a higher number of references per paper that other journals in the field.

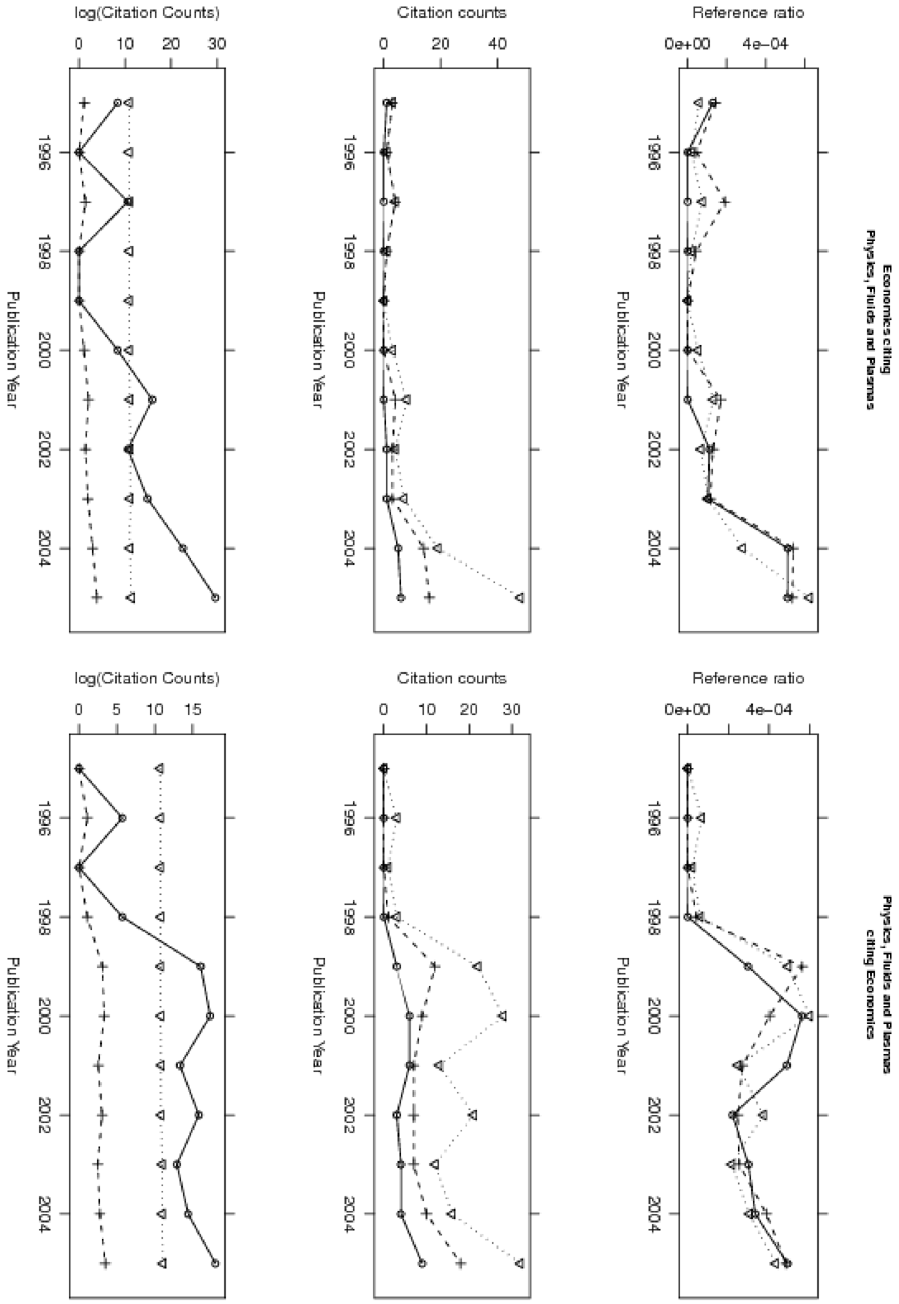


Figure 20 Example of a tableau of graphs used to analyse the growth characteristics of a selected pair and its reciprocal pair: the graphs in the top row show the citation share developments, those in the middle row the development of the absolute number of citations, while those on the bottom row show a comparison with the (endemic) growth in the originating fields, scaled to equal units .

pair was not considered. In order to find focus, publication content was assessed using a number of overviews, the most important of which were the following.

- A matrix of cited journals over years containing citation counts, as well one for citing journals over years.
- A list of best-cited publications, with title, journal, year of publication, and number of citations received.
- A list of citing publications that have the most citations to the cited field, with title, journal and publication year of both the citing publication and the cited publications.
- Lists of most active and best-cited organisations.

These overviews were also compared with those of the reciprocal pair.

8.3 Results and discussion

8.3.1 Search result

The distribution of the positive median growth rate values showed an approximately log- normal distribution with a mean of −2.43 and a standard deviation of 1.41. There were 683 pairs that had a value 1.5 times the standard deviation above the mean (−0.32). Applying the fast-growth and robustness requirements reduced this to number to 38 pairs. With the sole exception of the pair Biochemical Research Methods citing Statistics and Probability, none of the 38 pairs had any journals in common. **Table 10** lists all pairs. The first two columns of this table contain the names of the citing and the cited field. The column labelled RF deals with the research focus in the citing papers, and this column can contain four different values: F for a clear focus on a particular theme; P for a partial focus on a particular theme, with other minor themes also identified; G for a focus in a general, methodological sense (as opposed to a topical sense); and N if no focus could be found. In column R it is indicated whether the focus in the reciprocal (reversed) pair is the same as the focus in the pair found to converge: therefore, it only contains Y if a similar focus is found, and N if another focus is found or none at all[70] The other columns show indicators for the distribution of the citation counts: the total number of citations N, the median Med, the maximum Max, and the year of the maximum value in Peak.

One of the first things noticeable from **Table 10**, is the relatively small number of citations involved in the detected trends: an average of 251 and a median of 107 cita- tions. So, it appears that fields with higher number of citations between

---

[70] Note that if the Research Focus column contains N (no focus is found in the citing publications), then the value in the Reciprocal column can only be N as well.

them, do not show enough growth in order to be regarded significant, or do not meet some of the additional growth requirements. Figure ?? shows support for the first. This figure shows the logarithm of the median growth rate over the logarithm of the median size, for the pairs that have at least four observations. The solid diagonal line shown in this figure is a linear fit of the values, which (even though the fit is rather poor) illustrates the negative correlation between growth and size. Additionally, the dashed horizontal line indicates where the significant growth boundary of 1.5 times the standard deviation lies: only the points above that line were regarded in our search. The right-most point, corresponding to the largest median number of citations above that line corresponds to about 400. To explain this implicit limit, we again mention two possibilities. First, rapidly growing, field-surpassing developments are rather scarce, and developments are within the boundaries of a field. Second, the distribution may be dominated by smaller phenomena which show a relatively large growth-rate, although the growth in absolute number of citations is smaller.

8.3.2 Research focus

A summary of the type of focus in the result is given in Table 2, which leaves out the 12 pairs that showed no convergence. This table shows that in half of the cases a clear focus is found, while the other half is either accompanied by other topics, or is of a more abstract nature. However, this does not appear to negatively impact the presence of the reciprocal focus.

Focus	**Found**	**Reciprocal**
Focused	15	12
Partial	3	1
General	9	7
Total	27	20

Table 9 A summary of the type of focus in **Table 10**. The column Found contains the number of pairs, and Reciprocal contains the number of pairs that show the same focus in the reciprocal pairs.

8.3.3 Assessment of selected pairs

An exhaustive discussion of the content of the pairs listed in **Table 10** is beyond the scope this paper. Instead, we highlight pairs with interesting characteristics in an arbitrary order. The first example is the pair Economics citing Physics, Fluids and Plasma, which appears to be part of the larger “Econophysics” convergence (Stanley et al., 1999). The main source of the citations is the journal Quantitative Finance, and the citations are almost without exception going to Physical Review E. The best cited publication (counting only citations from Economics) is Plerou et al. (2002). On the reciprocal side, we find that most citations to Economics papers within the ten-year wide citation window are to publications from 2000 (as can be seen on the right half of **Figure 20**). This suggests that the developments in Physics took place before those in Economics, which is confirmed when inspecting the content of the citing publications from those years. Also, those Physics papers again refer to even older publications in Economics, the most cited of which is Arthur (1994). We therefore conclude that this area shows an area of mutual influence and exhibits an independent, reflective nature.

Another pair, Genetics and Heredity citing Communication, is an example of societal interest, and the influence of the creation of a new journal. The focus deals with the communication of consequences of research in Genetics, such as ethical consequences and risks. This focus is also present in the reciprocal pair. The journal responsible for most citations is New Genetics and Society, which started to be covered by the WoS in 2000. Since the top-cited Communication publications such as Kerr et al. (1998) were already dealing with this topic, as well as citing Genetics publications, we could infer that the newly created journal may have provided a more focused publication stage in Genetics and Heredity, moving the publications away from Communication, while continuing to refer to relevant publications there.

The pair Engineering, Industrial citing Agricultural Engineering mainly deals with topics related to Biodiesel, and shows a large Indian presence in the research. Also, judging from the titles of the cited publications, the research in the pair shows a transfer from basic science to applied science. Therefore, we consider it as an example illustrating economical, national interests. Interestingly, the reciprocal pair shows no connection to the research in the detected pair, but instead deals with miscellaneous applied agricultural topics. Therefore, according to our definition, this interdisciplinary development cannot be regarded as converging research.

The use of topics from the humanities by the natural sciences, is visible in the pair Physics, Nuclear citing Archaeology. The research has a partial focus (which means that also other, unrelated subjects were found) on the application of physics methods to archaeological artefacts. This is also found in the reciprocal pair. However, the application is not a new development, as the journal Archaeometry (which plays an important role in the reciprocal pair) was already established in 1958. Also, on further inspection, a single special issue of Nuclear Instruments and Methods in Physics Research B on “Radiation and Archaeometry” (N1-2, V226), turns out to be the most important reason for the selection of the pair. We doubt that a single special issue may be enough to label this example as converging research; instead, it may be an example of the import new tools from Physics, or alternatively, the export of specific problems to Physics.

As a last example we mention two related pairs: Computer Science, Theory and Methods citing Neuroimaging, and Optics citing Neuroimaging. Both pairs are part of the neuro-imaging and brain-imaging convergence that was also found in Buter et al. (2010a), but are representative of two different (related) themes: research into computational aspects of imaging, and research into optics applied to neuroimaging. The binding element is their cited knowledge base, because the top three cited journals is the same for both pairs: Neuroimage, Human Brain Mapping and

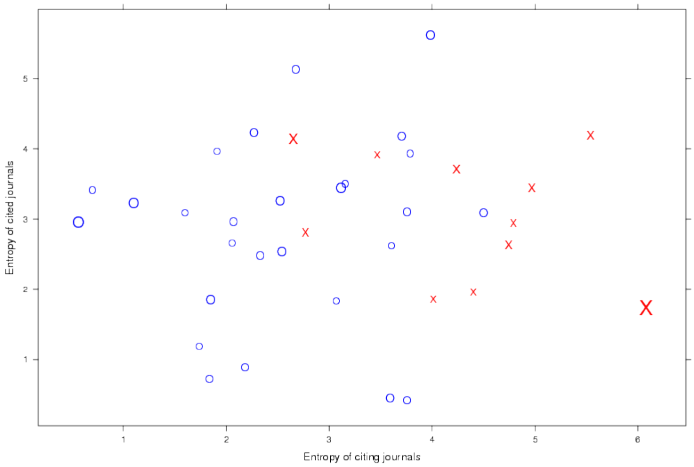


Figure 21 The entropy of citing and cited journals used as coordinates for the 38 resulting pairs, which are represented as a circle if they were found to have a research focus, and as a cross if not. The size of a circle or cross corresponds to the number of articles in a pair. On the left imaginary line running from (0,5) to (5,0) only pairs with a research focus appear, illustrating the weak correspondence between entropy and research focus

American Journal of Neuroradiology.

8.3.4 Useful elements in assessment

We found a number of elements more informative than others in establishing a research focus. The titles of the citing and cited publications were the most important sources of research topics, as well as the spread of these topics over cited and citing publications. Important indicators for the existence of focus were the sizes of the journal matrices: if such a matrix contained a lot of cited (or citing) journals, then focus was usually absent. Other overviews, such as those of citing and cited affiliations, or document types of cited and cited publication, turned out less useful in this respect.

To further explore the usefulness of journal matrices as indicators for focus, we quantified the spread of citations over journals by calculating the Shannon entropy[71], for both the cited and the citing journals matrix. These two numbers were used as coordinates in the scatterplot in **Figure 21**, where a circle indicates focus and a cross indicates no focus. Also, the size of the circle or cross is related to the total number of citations. From this figure we infer that there is a weak relation between

---

[71] e Shannon entropy is defined as H = -Ppi log(pi), where pi is the chance of event i, in our case the share of publications of a journal in the whole matrix.

the entropy values and the existence of a focus: below the diagonal that runs from (0,5) to (5,0), only focused pairs appear. Also, there appears no relation between the number of citations (size of a circle or cross) and the existence of focus. We consider this a useful first result and continue developing this indicator.

8.3.5 The nature of our results and converging research areas

The relatively small citation counts at the basis of our results challenges us to think about the nature of converging research. We hold that there are two important different devel- opments to discern. First there is the, possibly multidisciplinary, application of problems or tools. Such a development is typically short-lived, as the application does not lead to any new or deeper insights, and the research community looses interest. In the second type of development the community keeps interested, and the research starts to show some level of independence from its "mother"-disciplines, both at the cognitive level and the social level. When successful, this development will result in an interdisciplinary or even transdisciplinary research area (Van den Besselaar and Heimeriks, 2001). We consider the result of this second type of development representative for a converging research area.

Developments found by our search, will probably be the metaphorical tips of the ice- berg. To establish what we have detected, more information is needed about the larger scientific surroundings, and we may have to apply background knowledge, possibly even provided by experts. To interpret the larger scientific surroundings, we need to apply or even develop additional tools. At the social level, Berg and Wagner-Döbler (1996) hold that the structure of a research area can be seen as a combination of a "middle class" of authors around an established "prolific elite". The elite have an important say in which themes are considered important and subsequently provide opportunity to gain status or impact for members of the "middle class", as well as outsiders. Since citations also have a social dimension (Moed, 2005a), we therefore expect these structures to be visible in citation patterns, a fact which is also used in much of the work of Small (see e.g. Small and Upham (2009) for a recent example).

In the above description of example pairs, we have already established that some are indicative for larger, sustained research, like those representative of Neuroimaging and Econophysics. For the pair Genetics and Heredity citing Communication the nature is more difficult to assess, as it shows signs of an independent research area, since it has a spe- cialized journal; however, to confirm this we would have to inspect the research in a more detail. The pair, Physics, Nuclear citing Archaeology is also challenging in this sense, because even though we regard a special issue of a journal as the expression of an accumulated (and thus sustained) interest in a specific topic, without further investigation we cannot establish whether this special interest was the start of more research.

We can also note the following with respect to the relation between the research we located and "Mode 2" research in the sense of Gibbons et al. (1994). Mode 2 knowledge production is the ability of a network of practitioners to produce knowledge, while the codification of this knowledge is of lesser importance and may even be “part of the network”. Such knowledge is difficult to capture in bibliometric terms. At the same time, Mode 2 research requires a “context of application”, which may very well be related to the research focus we tried to establish in the different phenomena.

8.3.6 Conclusion and future research

We described a high-level, top-down methodology for searching convergence between fields using journal subject categories as a proxy for fields and citations as a measure for (interdisciplinary) application of research. A process was developed that consisted of two parts. In the first part, pairs of citing and cited fields were located using normalised citation counts as data and requirements with respect to growth, shape, and size. In the second part, the results were inspected, with the objective find research focus, i.e. the most specific common theme shared by most publications. We applied this process to WoS publication data for publications between 1995 and 2005. This resulted in 38 field pairs, which were inspected. After inspection, we found focus in 27 of the 38 pairs, and 20 pairs also showed similar focus in the reciprocal (reversed) pair. Also, interesting additional aspects were found in specific pairs, highlighting local, economical and societal interests, such as Biodiesel related research from India, or the ethical aspects of Genetics research.

Our search method has a number of clear advantages. First of all, it is data-driven, which makes it repeatable and applicable to new data, and even to other sources of data. Moreover, this makes the identification of converging research less dependent on the input of experts. Next, the method is easily adjustable as well as modular: search parameters such as the citation window can be tuned, and defining elements can be adapted. For instance, citation share could be calculated using the total number of citations obtained instead of given; or the journal subject categories could be replaced with another type of categorisation. Changing the required position of the peak may also be interesting, because a peak in the growth followed by enduring activity at a specific level may highlight developments that have managed to reach a steady state of activity.

Of course, the current implementation still depends on journal subject categories, which have well-known problems with respect to the delineation of related research. One problem is the coarseness, as there are only 243 categories to represent the whole of scientific research. One implication of the coarseness became apparent in the relation between size and growth shown in Figure ??, which illustrated that the larger a category becomes, the more difficult it gets to detect emerging developments between that category and others. Another problem is that developments between fields captured in a single category, cannot be detected. By employing finer grained categorisations we may address such problems, provided the categorisations still represent (mostly) different fields. Another problem of the JSCs is that some established (interdisciplinary) fields may be covered by multiple JSCs. However, we consider this less of a problem, because if significant changes are detected between related journals in different categories, there is still something potentially inter- esting going on.

Another aspect of the current implementation is the sampling of reference counts on a yearly basis, which is crude and introduces large variance in the time-series. Potentially, this could be replaced by a monthly sampling; however, this would introduce other problems, such as those related to quarterly appearing journals, and those which are for example published in "Winter 2005". Finally, the use of citations may present a problem of a more general, bibliometric nature: since most publications are published some time after the research was conducted, there is a delay between the use of the knowledge and the publication of that use, which could make it more difficult to detect trends that are in an early stage of development. Moreover, the time scales involved in the publication and citation processes differ in the various fields of science, as measured for instance by a field- specific age-distribution of references (Moed, 2005a). However, more research is needed before we can conclude what such effects have on the detection of interdisciplinary or (particularly) converging trends.

Future research will focus on improving the search method with respect to the points mentioned above. Additionally, the tools used to assess the results will be improved and extended, and parts of the assessment automated. The chart in Fig. 3 already provides some clues to how this automating could take place. Maps could be useful as well, such as cognitive science maps (Buter and Noyons, 2001, 2002a), or network maps (Calero et al., 2006 and Takeda and Kajikawa, 2009). As noted above, we also need tools which are able to provide indicators for the level of independence of the research taking place in a set of publications. Once such a level of independence is established, other interesting questions could be analysed, such as the development of "spill-over", i.e. the amount of knowledge that is developed in the convergence, but used in the originating, or other fields. A related question is whether a converged area, once it stops being active or attractive, would "dissolve" into the originating fields again, i.e. if the authors would return to their "mother"-disciplines, but keep on writing about the same issues ("annexation" by the originating field), or if the research into the topics completely stops. Finally, it may be interesting to see if some of the themes we found in the current application can be worked out in more depth and detail, not only with respect to the content of the research, but also to some additional bibliometric aspects, such as the development of the impact of converging research.

We hold that the value of our methodology lies in an interesting (apparent) paradox: scientific research is required to become more interdisciplinary in order to address complex societal, economical, technological and scientific problems, while at the same time researchers still tend to think in, organise in, and reward according to disciplinary lines. This tension provides a useful instrument, because research which does take the inter- disciplinary route, is implicitly taking that tension into account, and may therefore be onto something useful or interesting. We think that our methodology can provide the basis for identifying those "Jacks of many Trades" that take up challenges and may start new convergences in order to address complex problems.

References

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review, 84*(2), 406–411.

Berg, J. and Wagner-Döbler, R. (1996). A multidimensional analysis of scientific dynamics. *Scientometrics, 35*(3), 321–346.

Buter, R. K. and Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68.

Buter, R. K. and Noyons, E. C. M. (2002a). Using bibliometric maps to visualise term distribution in scientific papers. In *Information Visualisation, International Conference on*, pages 697. Los Alamitos, CA, USA.

Buter, R. K., Noyons, E. C. M. and Van Raan, A. F. J. (2010a). Identification of converging research areas using publication and citation data. *Research Evaluation, 19*(1), 19–27.

Calero, C., Buter, R. K., Valdés, C. C. and Noyons, E. C. M. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics, 66*(2), 365–276.

Carayol, N. and Thi, T. U. N. (2005). Why do academic scientists engage in interdisciplinary research?. *Research Evaluation, 14*(1), 70.

Cassirer, E. (1942). The influence of language upon the development of scientific thought. *The Journal of Philosophy, 39*(12), 309–327.

Davidse, R. and Van Raan, A. F. J. (1997). Out of particles: Impact of CERN, DESY and SLAC research to fields other than physics. *Scientometrics, 40*(2), 171–193.

De Boer, Y., De Gier, A., Verschuur, M. and De Wit, B. (2006). Building bridges: Researchers on their experiences with interdisciplinary research in the netherlands. Technical Report Den Haag: RMNO.

De Mey, M. (1982). *The Cognitive Paradigm.* London: D. Reidel Publishing Company.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Egghe, L. and Rao, I. K. R. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics, 25*(1), 5–46.

Gambardella, A. and Torrisi, S. (1998). Does technological convergence imply convergence in markets? evidence from the electronics industry. *Research Policy, 27*(5), 445–463.

Gibbons, M., Limogenes, C., Nowotny, H., Schwartzman, S., Scott, P. and Trow, M. (1994). *The new production of knowledge.* Sage Publications.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: the case of bioelectronics. *Scientometrics, 29*(3), 353–376.

Kerr, A., Cunningham-Burley, S. and Amos, A. (1998). Drawing the line: an analysis of lay people's discussions about the new genetics. *Public Understanding Of Science, 7*(2), 113–133.

Larivière, V. and Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126–131.

Moed, H. F. (2005a). *Citation Analysis in Research Evaluation*, volume 9 of *Information Science and Knowledge Management.* Dordrecht: Springer.

Morillo, F., Bordons, M. and Gómez, I. (2003). Interdisciplinarity in science: a tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology, 54*(13), 1237–1249.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, cocitation and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, 56*(12), 1250–1273.

Nordmann, A. (2004). Converging technologies — shaping the future of European societies. Technical Report EU HLEG Foresighting the New Technology Wave.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science, 50*(3), 242–253.

Persson, G. (1999). Mistra's role and some lessons learned in the first five years. *Policy Sciences, 32*(4), 323–326.

Plerou, V., Gopikrishnan, P., Gabaix, X. and Stanley, H. E. (2002). Quantifying stock-price response to demand fluctuations. *Physical Review E, 66*(2, Part 2).

Price, D. D. S. (1981b). The analysis of square matrices of scientometric transactions. *Scientometrics, 3*(1), 55–63.

Rinia, E., Van Leeuwen, T. N. and Van Raan, A. F. J. (2002). Impact measures of interdisciplinary research in physics. *Scientometrics, 53*(2), 241–248.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2001a). Citation delay in interdisciplinary knowledge exchange. *Scientometrics, 51*(1), 293–309.

Roco, M. C. (2002). Coherence and divergence of megatrends in science and engineering. *Journal of Nanoparticle Research, 4*(1), 9–19.

Roco, M. C. and Bainbridge, W. S., editors (2003). *Converging Technologies for Improving Human Performance.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Rosenberg, N. (1976). *Perspectives on technology.* Cambridge: Cambridge University Press.

Ruiz-Baños, R., Bailón-Moreno, R., Jiménez-Contreras, E. and Courtial, J. (1999). Structure and dynamics of scientific networks. part i: Fundamentals of the quantitative model of translation. *Scientometrics, 44*(2), 217–234.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics, 59*(3), 425–465.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to Non-Parametric regression curve fitting. *Journal Of The Royal Statistical Society Series B-Methodological, 47*(1), 1–52.

Small, H. E. and Upham, S. P. (2009). Citation structure of an emerging research area on the verge of application. *Scientometrics, 79*(2), 365–375.

Stanley, H. E., Amaral, L. A. N., Canning, D., Gopikrishnan, P., Lee, Y. and Liu, Y. (1999). Econophysics: Can physicists contribute to the science of economics?. *Physica A: Statistical Mechanics and its Applications, 269*(1), 156–169.

Takeda, Y. and Kajikawa, Y. (2009). Optics: a bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics, 78*(3), 543–558.

Upham, S. P. and Small, H. E. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics, 83*(1), 15–38.

Van den Besselaar, P. and Heimeriks, G. (2001). Disciplinary, multidisciplinary, interdisciplinary: Concepts and indicators. In *Paper the 8th ISSI Conference for Scientometrics and Informetrics, Sydney* .

Citing category	Cited category	RF	R	N	Med	Max	Peak
Medical Laboratory Technology	Agriculture, Multidisciplinary	F	Y	285	18	101	2005
Engineering, Industrial	Agriculture, Engineering	F	Y	147	8	64	2005
Ornithology	Agriculture, Multidisciplinary	F	Y	118	5	38	2004
Genetics and Heredity	Communication	F	Y	117	6	42	2005
Immunology	Construction and Building Technology	F	Y	111	6	27	2005
Economics	Physics, Fluids and Plasmas	F	Y	98	4	48	2005
Computer Science, Software Engineering	Food Science and Technology	F	Y	81	5	27	2004
Agriculture	Soil Science,Physics,Fluids and Plasmas	F	Y	78	4	30	2005
Physics, Applied	Agriculture, Multidisciplinary	F	Y	70	3	29	2004
Materials Science, Characterization and Testing	Dentistry, Oral Surgery and Medicine	F	Y	61	2	30	2005
Imaging Science and Photographic Technology	Engineering, Geological	F	Y	47	2	31	2005
Social Sciences, Mathematical Methods	Biotechnology and Applied Microbiology	F	N	195	10	58	2005
Physics, Particles and Fields	Computer Science, Information Systems	F	N	121	4	68	2005
Remote Sensing	Engineering, Geological	F	N	55	2	36	2005
Biochemical Research Methods	Statistics and Probability	G	Y	4480	127	2111	2005
Computer Science, Theory and Methods	Neuroimaging	G	Y	646	9	248	2004
Cell Biology	Nanoscience and Nanotechnology	G	Y	452	23	164	2005
Biochemistry and Molecular Biology,	Materials Science, Composites	G	Y	129	10	53	2005
Physics, Nuclear	Archaeology	G	Y	83	6	32	2004

Table 10.a The 38 pairs fitting all search requirements. The RF column indicates the research focus, which can be either F, P, or G. The R column indicates whether or not the focus in the reciprocal pair is the same as one found in the listed pair. N contains the total number of citations between 1995 and 2005 using a 10-year wide citation window. The next columns deal with the distribution of the citations: the median and maximum value, and the year of the maximum value in the peak column.

Citing category	Cited category	RF	R	N	Med	Max	Peak
Geosciences, Multidisciplinary	Parasitology	G	Y	72	3	28	2005
Horticulture	Allergy	G	Y	51	3	26	2005
Optics	Neuroimaging	G	N	208	12	92	2005
Engineering, Electrical and Electronic	Chemistry, Medicinal	G	N	137	9	33	2004
Physics, Fluids and Plasmas	Chemistry, Applied	P	N	187	9	73	2005
Telecommunications	Energy and Fuels	P	N	103	7	32	2005
Ornithology	Biophysics	P	N	67	2	29	2005
Materials Science, Composites	Biochemistry and Molecular Biology	N	N	266	7	155	2005
Engineering, Manufacturing	Surgery	N	N	171	11	58	2005
Automation and Control Systems	Cell Biology	N	N	166	9	55	2004
Computer Science, Theory and Methods	Urology and Nephrology	N	N	117	2	61	2004
Medicine, Research and Experimental	Applied Linguistics	N	N	112	6	36	2005
Engineering, Manufacturing	Cardiac and Cardiovascular Systems	N	N	97	4	29	2005
Materials Science, Ceramics	Microbiology	N	N	86	3	42	2005
Water Resources	Engineering, Biomedical	N	N	78	5	31	2005
Film, Radio, Television	Psychology, Experimental	N	N	68	2	29	2004
Cardiac and Cardiovascular Systems	Integrative and Complementary Medicine	N	N	67	3	27	2005
Computer Science, Information Systems	Nanoscience and Nanotechnology	N	N	66	2	30	2005
International Relations	Psychiatry	N	N	65	4	33	2005

Table 10.b The 38 pairs fitting all search requirements. The RF column indicates the research focus, which can be either F, P, or G. The R column indicates whether or not the focus in the reciprocal pair is the same as one found in the listed pair. N contains the total number of citations between 1995 and 2005 using a 10-year wide citation window. The next columns deal with the distribution of the citations: the median and maximum value, and the year of the maximum value in the peak column.

9

Identification and analysis of the highly cited knowledge base of sustainability science*

Abstract

At its hart, Sustainable Development (SD) is a concept that expresses a desirable, yet abstract goal. And as a consequence it has started to cover a broad set of meanings. Another consequence is that this makes it a challenge to unambiguously relate scientific research publications to SD. Even more, sustainability goals may be supported by scientific research that did not have this specific goal in mind. This is particularly true for the referenced scientific literature of these 'sustainability publications', i.e. the scientific knowledge base that represents the roots of sustainability science. In order to study this knowledge base, we combine a lexical search to find relevant publications with a search for the most cited publications. We assess the result by: 1) looking at the science fields in which the publications are published; 2) a map of publication clusters; and 3) the themes associated with the publications in these clusters. Moreover, we contrast this with other sets of publications relevant to SD in order to find the research orientation of these sets. We find that our approach supports the division of publications over the three pillars of SD: environment, economy, and sociology. However, these pillars are not equally represented in peer-reviewed journal publications. In addition, we find that the more specialised 'sustainability journals' in the Web of Science database do not represent the breadth of the whole field of sustainability science. Finally, we do not yet see evidence that the field of sustainability science is moving into a more transdisciplinary (integrated) state.

---

* Submitted to 'Sustainability Science'.

9.1 Introduction

The concept ‘Sustainable Development’ (SD) is not a crisply defined, scientific concept, and it even ‘transcends’ science in the sense of Weinberg (1972). Still, to reach sustainable goals, it is clear that a firm base in science is important. Since the end of the twentieth century, efforts have been under way to create an integrated “sustainability science” (Dodds, 1997; Kates et al., 2001; Clark and Dickson, 2003 and Clark, 2007). Such efforts include initiatives like Komiyama and Takeuchi (2006) and Keitsch (2010) to create a framework that should encompass and stimulate research in the diverse aspects of SD. Of course, even without the support of such frameworks, much research had already been done that dealt with topics related to sustainability. To identify such research is interesting as it allows us to see where this research originated and what it addressed so we can to find and analyse the actors, measure their relative success, unravel their interactions, and to monitor changes and developments. The information we could use as a compass (Lee, 1993), but also to support SD initiatives in a policy context, for example to find experts in particular subjects, or organisations within a network of excellence.

The diversity of SD has been noted by, among others, Phillis and Andriantiatsaholiniaina (2001), Mayer (2008), Kates et al. (2005) and Kajikawa et al. (2007). Already some twenty years ago Holmberg and Sandbrook (1992) noted that they were aware of *seventy* different definitions. Another illustrative example is the multitude and diversity of the ten parts within a single volume by Kirkby et al. (1995): Biodiversity; Climate Change and Energy; Population; Agriculture; Industrialization and Pollution; Urbanization and Health; The Commons; Environmental Security and Environmental Institutions; Empowerment; and Environmental Economics. Such breadth even caused some authors to fear that SD risked becoming meaningless (Sharachchandra, 1991).

This multifaceted nature makes it already difficult to operationalise the concept, especially in terms of scientific publications. However, even if we could find a satisfactory operationalisation, at what point would we consider scientific research related to SD? This important issue is probably best illustrated by an example. Take research into new building materials. Such materials could be lighter, yet isolate better, and require less toxic materials to create. Clearly, such materials would be a good fit for sustainable urban development projects. However, should the research into those materials be labelled as research into SD and its applications? The basic research underlying these materials was not done with the (specific) intention to use it in any SD project. Now, Clark and Dickson (2003) points out that many interactions between science and SD are problem driven (*ibid*, p.8059). So, one way to draw the line between publications directly supporting SD and those indirectly supporting SD, is by finding out whether the authors intended to address a problem related to SD. But then, we are faced with the problem of finding out what the authors’ intentions were and if the problems they addressed had an SD focus. Alas, we cannot interview all authors, but we can look for papers which have been labelled by their authors as related to SD, simply by looking for publications which have in their title a word which starts with 'sustainab'. Since a title is more purposeful instrument (Buter and Van Raan, 2011; Soler, 2007 and Noyons, 1999), we consider the occurrence of such words an even stronger signal than the occurrence in an abstract.

After we have identified these publications, we can then look 'underneath' this set, by collecting the cited publications. The role of an individual reference from one scientific publication to another may vary: some are included with social or political motives, others to indicate a relation based on content. A nice overview of the different opinions on the role of references and citations is given by (Moed, 2005a, p.194). Still, for highly cited publications, Peters et al. (1995) showed that not only is there a link between the (highly) cited publication and the citing publications, but *also* that the citing publications are content-wise related to each other if they share references to such a highly cited publications. As a result, by using highly cited publications, we have a good chance of finding publications which share related research topics. Still, the context in which the highly cited publications have been written, may be different from the context in which the citing publications are written. In the past, we have used this phenomenon to locate converging research by searching for fast-growing citation patterns from one field to another (Buter et al., 2010b). Given the multi-faceted nature of SD, this is one of the things we can expect.

Many authors have expanded on the diverse challenges within SD. Some have taken a systemic approach from an educational perspective, like Elliot (2006). Others have taken a specific part of SD, like environment and economics, and expanded on that (Pearce et al., 1990). With respect to the interactions between SD and science, many have taken an abstract, qualitative approach. Komiyama and Takeuchi (2006) for example, used their extensive experience to differentiate between three levels ('global', 'social', and 'human') and use this to classify problems in sustainability, not only at these three levels, but also in the interaction between these levels. Clark and Dickson (2003) use their knowledge of the field to emphasise the pragmatic character of sustainability science, the interaction between scholars and practitioners, and the associated challenges and complexities. Kates et al. (2001) stresses the interaction between nature and society and proposes a set of research questions which can be addressed by sustainability science. Kajikawa (2008) holds an interesting middle ground, and combines a qualitative and quantitative analysis to arrive at a research framework for sustainability science. Our paper features such a quantitatively driven approach. We used a well-tested combination of citation analysis and topical analysis, using clustering and mapping. Such an approach was pioneered by Braam et al. (1991a), but was also used by for example Noyons et al. (1999). Therefore, our approach also closely resembles that of Kajikawa et al. (2007) who used it to analyse publications that contain words starting with 'sustainab' in their title or abstract. As such, we can use their results to compare and contrast with the results we present in this publication. Nevertheless, our method differs from their approach on important points, as will be explained in the next section. In addition, our focus is on the analysis of the highly cited knowledge base and to contrast its origins and topics with those of other publications sets dealing with SD related research.

9.2 Data and Method

The basis for our analyses is the Web of Science (WOS)[72], a bibliographic database which covers a large part of the global, scientific output published in peer-reviewed journals. Our institute (CWTS, Leiden University) used the WOS to create a system that is fully dedicated to bibliometric analyses and indicator calculations. The database contains important elements of publications, such as title, abstract, affiliations, keywords, or journal, as well as the references to other publications, many of which are also part of the WOS.

We already outlined our approach for finding SD related publications above, but for the sake of clarity, we discuss the options we have at our disposal to operationalise or delineate research related to SD. One option is to start with *journals* that have SD in their scope, which we can recognise because the word 'sustainable' or 'sustainability' is in their title. Currently, the WOS covers twelve such journals, with 3, 556 publications. The journal with the longest coverage is the *Journal of Sustainable Agriculture*, which has been covered starting from 1991 and has 970 publications in the database. Other journals started later, mostly 1999 and onward[73]. The advantage of using such a set of journals and the publications covered by these journals is that they are readily available and are clearly related, because of the peer-reviewed nature of these journals. However, the most important disadvantage is that important contributions in other journals (most notably from prestigious, multidisciplinary outlets such as *Science*, *Nature*, and *PNAS*) are not covered. So, an alternative option would be to create a thematic filter, which contains expressions (words, phrases and their combinations) related to SD, and to collect publications which match this filter. The advantage is that publications from the whole database are can be selected, also those in the prestigious journals. The disadvantage is that is rather difficult to create such a filter, because it requires the input and consensus from experts, potentially from different fields. A third option is to start with important contributions to the field, which can be found using a citation analysis or (again) by consulting experts. Subsequently, the publications that cite these contributions can be collected. To extend this selection, publications which share a lot of references with these citing publications can be added as well, though it is not straightforward what to include and what not. The advantage is that it is usually rather straightforward to collect a small number of important contributions. However, the disadvantage is that, again, it can be difficult to be complete or to reach consensus on what should be considered an important contribution and what not.

In this paper we built on the idea that the label 'sustainable' or 'sustainability' is a strong indicator for the intended purpose of a publication. Hence, we started with a straightforward filter that only matches the expression 'sustainab*' in the titles of publications published in the WOS between 1999 and 2008. The purpose of this

---

[72] The WOS database has been licensed by CWTS from its producer, Thomson Scientific.

[73] Kajikawa et al. (2007) give a complete listing of these titles.

first step was to collect publications that have been labelled *explicitly* as dealing with sustainability or sustainable development. The resulting set of publications we will refer to as the *seed set.* Then, we collected publications *cited* by the seed set. In bibliometric terms, this second set is called a knowledge base (Braam et al., 1991a). A complete knowledge base (KB) is much larger than set of citing publications, and many publications are cited only once or twice. To reduce this sparsity of the KB, we required a minimum number of citations a (KB) publication should receive, first from the seed set, but then also from the larger environment (the rest of the WOS). This differs from the approach taken by Kajikawa et al. (2007) who focused on the maximum connected component within the seed set. This is a graph or (social) network concept, which expresses the largest set of publications that are linked by citing one another. The advantage of our approach is that we can extract different, unrelated clusters of cited publications, which are not necessarily linked by citations, while still arrive at a set of publications for which consensus exists that it has been relevant or representative for research in the seed set.

The larger environment of sustainability research includes papers from many fields. Since there are large differences in the length of the reference lists in different fields, there are also large differences in the expected number of citations for publications in these fields. As a result, we cannot use an *absolute* minimum number of citations to express highly cited. Therefore, we should use a *field-normalised* citation indicator. We chose the field-normalised citation measure CPP/FCSm, which was developed and used by CWTS in most of its bibliometric studies. For the calculation of this indicator, the WOS journal subject categories (JSCs) are used as proxies for fields (Moed et al., 1995 and Van Raan, 1996). We calculated CPP/FCSm values for the period 1999-2008. A value of 1.0 for the CPP/FCSm means that a publication is cited along the average of the field. Therefore, publications that have a value larger than 1.0 are cited above average. However, what value can be considered as the threshold between well cited and highly cited, is a more difficult question. In the past, we have experimented with the distribution function of CPP/FCSm values within a field (Buter et al., 2010a), but it remains difficult to arrive at a completely objective approach. Therefore, we used our experience and some experimentation to settle on the following: at least five citations from the seed set and a CPP/FCSm > 1.5[74]. The thus identified publications we consider as members of the highly cited knowledge base (HCKB).

We investigated the interdisciplinary character the HCKB by looking at the science fields from which the publications originate. As noted above, we use JSCs as proxies for science fields. JSCs have well-known problems with respect to their delineation of related research, and provide a rather high-level[75] classification to represent the

---

[74] Since a CPP/FCSm value of 1.0 means that the impact is precisely the field-average, a value of 1.5indicates a 50% higher number of citations than the average. Based on our experience with performing bibliometric analyses, we know that a value of 1.5 can be used to identify highly cited publications.

[75] At the time we conducted our analyses, there were 246 JSCs.

whole of science. Nevertheless, they are a readily available categorisation of related publications, their definition is standardised in a transparent way, they are quite stable over time, and their use is consistent with our use of the CPP/FCSm.

Although we chose not to use the journals that have 'sustainable' or 'sustainability' in their name as the basis for our analyses, they still provide a valuable source of information, which we used to evaluate the coverage of the HCKB. Since the number of these journals is small (twelve), they cover only a small (eight) number of fields. Again, it is therefore more informative to use the KB of these publications and to compare the coverage of that KB with the HCKB. To mediate the sparsity of this KB, we required at least two citing publications from the journals. Since this already reduced the number of KB publications considerably, we did not require an additional minimum value for the impact. We refer to the resulting set as the journals' knowledge base (JKB). Finally, another relevant set is the one containing all[76] publications citing the HCKB, which can be considered the larger research environment of the publications in the seed set. **Figure 22** provides a schematic overview of the different publication sets we use in this paper. The size of the ellipses do not precisely represent the size of the sets, we merely wanted to express how the sets overlap. The outer set is the whole database (WOS), all other sets are a subset of the WOS. The complete KB of the seed ('Seed KB' in the figure), partially overlaps with the seed set itself, because there are highly cited publications in the seed set as well. Additionally, the seed set overlaps with the set of publications from the journals ('Journals' in the figure), as well as the journals' knowledge base ('Journal KB'). The figure also shows that the HCKB is a complete subset of the whole KB of the seed, and partially overlaps with the JKB.

We investigated the structure of the HCKB by clustering its publications. The clustering algorithm we used was recently developed at CWTS (Waltman et al., 2010) and can be described as a more general version of the Girvan-Newman community clustering algorithm (Girvan and Newman, 2002). The input for the clustering is the sum of the co-citation and bibliographic-coupling matrices. A co-citation matrix is a symmetric matrix that captures the number of times two publications are cited in the same reference list of a citing publication. A bibliographical coupling matrix is also a symmetric matrix, but captures the number of times two publications cite the same cited publications. The sum of the two matrices does not convey any particular phenomenon, but because the publications on the rows and column of the matrices are the highly cited publications from the HCKB, it is a simple enforcement of the correspondence between two publications. The output of the algorithm is a unique mapping between a publication and a cluster. A user can influence the number of clusters indirectly, by filling in a non-default value for the 'resolution' parameter: a higher value implies a higher number of clusters, which subsequently have a smaller size. However, the precise number of clusters is determined by the clustering algorithm automatically. From the resulting clustering, we also created a two-dimensional map. For this, we again had several options, such as MDS (used in,

---

[76] Published between 1999 and 2008, but not necessarily in the seed set.

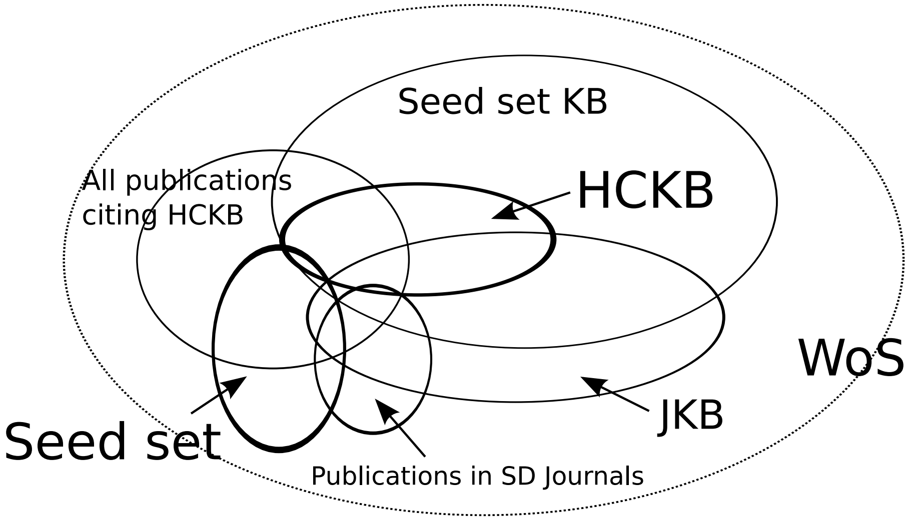


Figure 22 A schematic overview of the different sets of publications used in the article: the seed set, the KB of this set (wavy pattern), the HCKB (dark vertical stripes), the set of publications from the 'sustainable' journals, the KB from this set (JKB, small spots), and the set of all publications citing the HCKB (large spots).

for example, Buter and Noyons (2002a)) or a network-related visualisations such as Kamada and Kawai (1989). Instead, we chose for the VOS method, described in Eck and Waltman (2009). The VOS method has been designed with the aim to enhance the graphical representation of bibliometric maps in order to aid the interpretation of these maps.

To compare the orientation of research in the HCKB with that in the other sets, we used both fields and clusters, as they capture the research environment of the publications and the research topics in these publications. We ranked the fields and clusters according to their coverage by the HCKB and plotted the share of papers as a graph (see **Figure 24**, below). This approach was used in Noyons and van Raan (1998) to compare the shift in cognitive orientation in the research area of 'Optimisation'. A spike in this graph implies that there is an orientation on a specific field or cluster in one set that is not present in to base set (the HCKB, in our case), while a valley implies that the base set orients on research not represented in the other sets.

To assess the content of the publications in the clusters, we used frequently occurring noun phrases (NPs) from titles and abstracts. However, not all frequent phrases are useful. A general phrase like 'paper' occurs frequently in many publications but does not provide much information. Still, even specific phrases are not always useful, because they could be considered as common jargon in a research area and used in many different publications in that area. As a result, such specific phrases would occur in almost every publication cluster, and we cannot use it to identify the research topics specifically addressed in that cluster. To filter such phrases, we used the following approach. First, we applied a list of 'non-informative' phrases[77], which we have developed over the years, and includes phrases such as 'paper','result', and 'improved method'. Then for the remaining NPs, the relative occurrence over all clusters was calculated and subsequently used to calculate a Shannon entropy value: $\sum p_i \ln(p_i)$. If an entropy value is high compared to the values of other phrases, then the phrase provided little information and was discarded. This approach is similar to the one used in Quinlan (1993), Buter and Noyons (2005) and Buter et al. (2010a). Although the list of non-informative phrases introduces a slight subjective bias, our approach is quite similar to the NC-values calculated by Kajikawa et al. (2007), which also includes the Shannon entropy as one of its ingredients. However, we do not use NPs as an additional source for clustering, but only to support the assessment of the resulting clusters. The resulting list of NPs per cluster gives a first indication of the research topics in a cluster and were used in combination with the names of frequently occurring journals and fields, to devise labels for the clusters.

Since we cluster the HCKB, an interesting question is whether the publications in the clusters are cited from different fields, and if this changes over time. In a sense, this measures the diversity of the research in the clusters (Rafols and Meyer, 2009). We measure this again by calculating the Shannon entropy over the occurrence of fields in a cluster. The occurrences are divided by the total number of occurrences in the cluster j: $p_i = n_{ij}/\bar{n}_i, \bar{n}_i = \sum n_{ij}$. Lower values imply lower diversity, i.e. the citations come from fewer fields. We would expect a research are that shows signs of becoming transdisciplinary, to show an increasingly lower diversity value, because the research it draws upon becomes increasingly self-supporting as well as becoming more specialised.

9.3 Results and Discussion

9.3.1 General

We start with some general results. In **Table 11** we show the number of (WOS covered) publications (P) in our seed set, broken down per year together with the indexed growth (1999 = 100). In total, the seed set contains 10,594 publications. The table shows that the number of publications that have 'sustainab*' in their title is steadily growing: from 674 publications in 1999 to nearly triple that amount (1,835 publications, 272%) in 2008. Of course, these numbers are lower than the number of publications reported in Kajikawa et al. (2007), but the fast growth is similar.

The publications in the seed set, cite 66,294 different publications. As noted above, many cited publications are cited only once or twice by seed set: 55,013 are cited only

---

[77] This is a list composed by us on the basis of the results of many CWTS studies with NPs from scientific publications; it can be requested for research purposes from the first author of this publication.

Year	P	Growth
1999	674	100
2000	789	117
2001	751	111
2002	787	117
2003	946	140
2004	965	143
2005	1,192	177
2006	1,245	185
2007	1,410	209
2008	1,835	272
Total	10,594	-

Table 11 The number of publications P per year containing a word starting with 'sustainab' in their title or abstract and the indexed growth (1999=100).

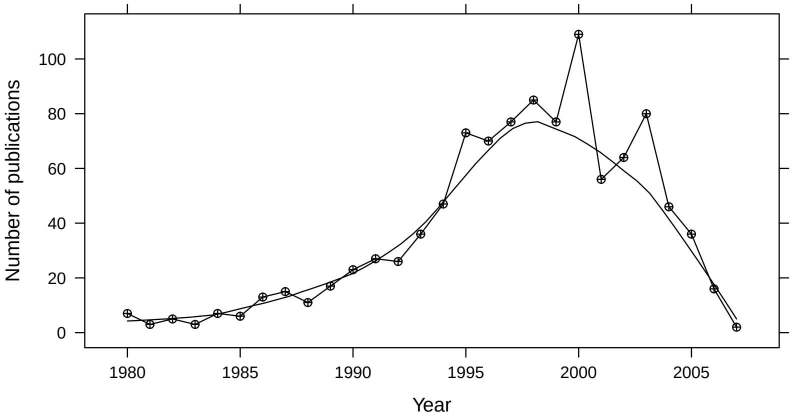


Figure 23 The number of HCKB publications per year.

once, 11,281 publications are cited twice, and only 1,379 are cited by five or more[78]. After calculating the field-normalised impact of the publications, we found that 1,038 were cited by more than five seed publications have a CPP/FCSm value above 1.5. As noted above, we call this set the *Highly Cited Knowledge Base* (HCKB). We did not restrict the publications in the HCKB to the period 1999-2008, and **Figure 23** shows the number of publications broken down by year. Most cited publications were published between 1994 and 2004, which is not surprising given that authors prefer to cite recent work and the seed set contains publications from the period 1999-2008.

---

[78] Note that we only indicate the number of citations from the seed set; the number of citations from all WOS covered publications could be larger.

The sharp decline between 2000-2001 should probably be considered as a database artefact and we do not attach any significant meaning to it; the smoothed version of the distribution is therefore probably a more accurate description.

To see how coherent the HCKB is, we look at how many seed publications *share* a reference to a publication in the HCKB[79] (see also Rafols and Meyer (2009)). As it turns out, 29% (3,063 out of 10,594) of the publications in the seed set share one or more references to the HCKB publications. This percentage is comparable to the percentage of papers in the maximum connected component of Kajikawa et al. (2007). Moreover, if we take into account that the seed set is for only 42% (4,484 out of 10,594) coupled on the *whole* KB (which contains 66,294 publications), we see that we started out with a very loosely coupled set of publications in the first place. Thus, alternatively, the HCKB can be regarded as representative for over two thirds (68%, 3,063 out of 4,484) of the bibliographically coupled publications in the seed set.

9.3.2 Fields

Table 12 shows the prevalent fields (in terms of the JSCs of the WOS) in the HCKB, and the corresponding number of papers for those fields in the other sets. At the top, we find fields related to economical and environmental research. Further down, we also find agriculture, chemistry, and (ranked thirteenth) sociology.

To compare the differences in coverage of these sets more effectively, we created **Figure 24**, which shows the share of the fields in the different sets in the same order as in **Table 12**. We have added labels to some of the most important outliers (spikes and valleys). We immediately observe the large spikes for the JKB in the fields *Energy & Fuels*, *Agronomy*, *Soil Science*, and *Plant Sciences*, which means that these fields are represented much more in the JKB than in the HCKB. Most likely, these spikes are the result of the bias introduced by the specialised journals in this set, the *Journal of Sustainable Agriculture*, and the journal of *Renewable & Sustainable Energy Reviews.* The valley labelled *Economics* means that this research is more specific to the HCKB, compared to the other sets; interestingly, this also includes the seed set used to create the HCKB. *Ecology* on the other hand has a higher relative coverage(about 15%) in the set of all publications citing the HCKB (the all-citing set), and shows an interesting spread of coverage in the other sets, notably the seed set with a coverage of about 7%. The fourth field, *Environmental Studies* is covered better than *Ecology* in the seed set, but much worse in the all-citing set. The *Multidisciplinary Sciences*, the field in which journals like *Science* and *Nature* reside, have (as could be expected) a higher coverage in the HCKB than in the seed set and in the all-citing set. However, also the JKB shows rather low coverage of publications in this field, which could be related to the high coverage of the other fields noted above. *Management* is another field that is covered more in the all-citing set, but much less in the seed set. Finally, *Engineering, Environmental*, is covered rather well by the seed set.

---

[79] In bibliometric terms this is called *bibliographic coupling*

#	Field	$\mathbf{n}_{\mathrm{HCKB}}$	$\mathbf{n}_{\mathbf{seed}}$	$\mathbf{n}_{\mathbf{all\ citing}}$	$\mathbf{n}_{\mathrm{JKB}}$
1	Environmental Sciences	228	2,003	12,523	435
2	Economics	200	757	9,808	160
3	Ecology	194	890	15,336	300
4	Environmental Studies	183	1,400	5,579	243
5	Agriculture Multidisciplinary	86	513	2,000	202
6	Multidisciplinary Sciences	86	254	1,721	69
7	Planning & Development	82	621	2,918	83
8	Management	79	271	8,781	34
9	Business	75	174	6,629	15
10	Geography	40	528	2,564	32
11	Energy & Fuels	40	453	1,620	634
12	Engineering Environmental	34	722	1,937	56
13	Sociology	33	203	2,716	62
14	Forestry	30	397	2,757	15
15	Biodiversity Conservation	25	77	3,253	18
16	Urban Studies	22	357	1,405	23
17	Agronomy	22	252	2,204	332
18	Fisheries	18	120	1,876	0
19	Engineering Chemical	17	327	1,937	85
20	Marine & Freshwater Biology	17	164	4,231	3
21	Social Sciences Mathematical Methods	16	17	1,027	2
22	Soil Science	16	124	3,436	236
23	Water Resources	15	569	1,856	52
24	Operations Research & Management Science	15	91	1,686	21
25	Psychology Multidisciplinary	14	46	2,263	7
26	Social Sciences Interdisciplinary	13	100	1,135	7
27	Statistics & Probability	12	9	604	1
28	Chemistry Multidisciplinary	12	229	2,740	8
29	Public Environmental & Occupational Health	12	226	2,364	2
30	Plant Sciences	12	141	4,626	127
31	Construction & Building Technology	11	184	184	15

Table 12 The most prevalent science fields in the HCKB (at least 10 publications), together with the corresponding number of publications in the seed set, the set of all publications citing the HCKB, and the JKB.

From these findings we conclude that there appears to be a particular orientation in the seed set publications that is different from the all-citing set of publications, although they share the same KB. We see that the all-citing is more focused on *Ecology, Management, Marine & Freshwater Biology*, and *Plant Sciences*, whereas the seed set has a focus on *Environmental Sciences, Environmental Studies, Environmental Engineering*, and *Water Resources.* This affirms that the seed set introduces a bias towards environmental sciences. It could be possible that this bias was intended, and a word like ‘sustainable’ or ‘sustainability’ fits the vocabulary of these fields better

The coverage of fields is quite comparable in the top 30 of the different sets. How-

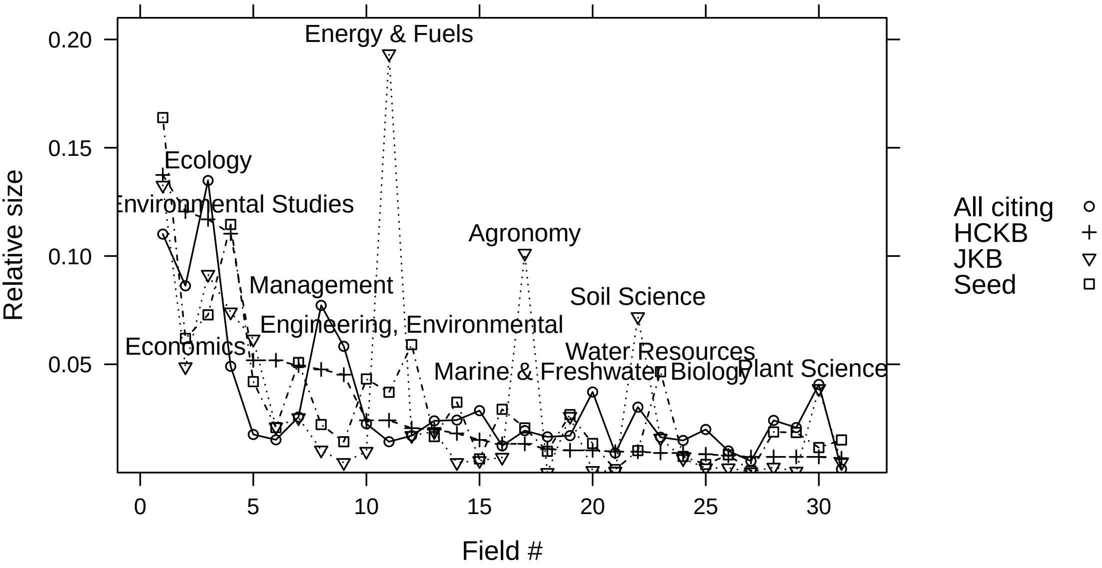


Figure 24 Relative publication counts from **Table 12**, same rank numbers(#) as in **Table 12**. Some important outliers have been labeled with the field to which they belong.

ever, some fields are not in the top most common fields in the HCKB, but are covered quite well by other sets. The fields *Thermodynamics* (#9 with 149 publications) and *Mechanics* (#11 with 92 publications) are covered quite well by the JKB. Also, we find *Psychology* ranked 14$^{\text{th}}$ in the all-citing set, with 2719 publications; however, it has no prominent coverage in the other sets.

9.3.3 Clusters

We move on to the analysis of the clusters in this set of publications. We found that the HCKB could be divided into 13 clusters, which are shown in **Table 13**. The table shows the number of publications that make up the cluster (P), the number of citations from the seed set, from the all-citing set, and from the JKB. Additionally, we have included the overlap of the publications in the cluster and the publications in the JKB (O_{JKB}), which captures where the two KBs overlap and indicates what topics are shared. The last column contains the most informative NPs. We extracted these NPs from the titles and abstracts of the clustered papers with at least five associated publications. Using the methodology explained above, we selected only those NPs that can be considered as most informative. Clusters 11 and 13 did not yield any NPs that were both significant and had enough associated papers. For these clusters, we inspected the list and manually selected the NPs shown in **Table 14**. Although

#	M	Label	P	%	C_{seed}	C_{all}	C_{JKB}	O_{JKB}	NPs
1	1	Ecology, economics and impact.	540	52	1,958	2,1521	462	195	energy, environmental impact, economy, science
2	1	Forest ecology	150	14	564	23,680	104	27	forest, conservation, landscape
3	1	Agriculture	62	6	2 56	16,655	164	31	nitrogen, soil quality, diversity
4	2	Production, business	49	5	199	5,664	21	10	business, environmental management, environmental performance
5	1	Fishery, aquatic science	45	4	201	4,872	28	1	fishery, marine reserve, marine ecosystem
6	1	Sustainable grazing system	37	4	54	700	1	0	pasture, sustainable grazing system, persistence
7	2	Business, management	32	3	95	12,005	15	1	social capital, strategic management
8	2	Production, management	29	3	112	8,205	19	6	transition, ionic liquid
9	3	Econometrics	29	3	63	4,706	2	0	cointegration, unit root, deficit
10	4	Sociological aspects	28	3	92	14,068	15	4	behavior, attitude, belief
11	2	Health projects	17	2	53	612	0	0	intervention, health promotion intervention
12	3	Politics, economy	16	2	85	3,631	10	2	economic growth, economic development
13	5	Polymer science	3	0	12	0	0	0	resin, hydroxythel
Total			1,037	100	3,744	116,319	841	277	

Table 13 Clusters of the HCKB, the absolute and relative number of publications P, as well as the number of citations from the seed set, the all-citing set C_{all}, and the JKB C_{JKB}. Also given in the overlap O_{JKB} with the JKB. The first column M indicates the meta-cluster in which the cluster is placed and has been used to sort the table.

the number of papers associated in the clusters with particular NPs rather quickly becomes very low, this table nicely illustrates how the concepts in the clusters differ from cluster to cluster. It also shows that the 'Ecology, economics and impact' cluster contains a mix of publications dealing with diverse topics. The descriptive label in the last column is based on the NPs in the last column.

The table shows that the number of publications in the *Ecology, economics and impact* cluster is high compared to the other clusters, as over fifty percent of the publications are in this cluster, making this necessarily a general cluster. Indeed, its more prevalent journals are *Ecological Economics* (70 publications) and *Agriculture Ecosystems & Environment* (21 publications). The next is *Forest Ecology* with *Conservation Biology* as its most prevalent journal (17 publications), and *Agriculture* with *Nature* (15 publications). The next cluster *Production, business* is in another meta-cluster and has the *Academy of Management Review* as its largest journal (13 publications).

The clusters can also be displayed in a two-dimensional map, which is shown in

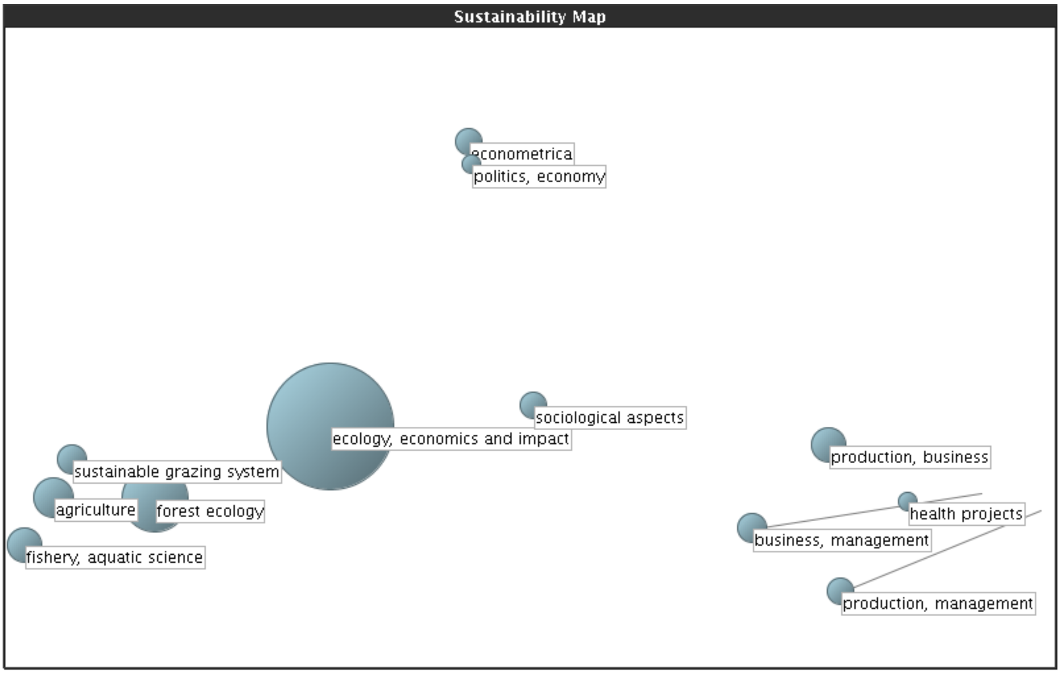


Figure 25 The map of the HCKB. The size of the clusters is relative to the number of publications covered by the cluster. Some of the clusters have been moved in order to include their labels in the map, and lines are drawn that connect them to their original position.

Figure 25. The map shows a clear distinction in three parts, corresponding to the meta-clusters 1, 2 and 3. The fourth meta-cluster contains solely *Sociological Aspects* and is located between clusters 1 and 2. The cluster *Polymer Science* had too few papers to give it a reliable position on the map.

The accompanying **Figure 26** shows the 'spectral' analysis, similar to **Figure 24**. This figure highlights a number of interesting findings. First, it shows an almost perfect similarity between the distribution of papers over the clusters of HCKB and the seed set. This has to be expected, since the seed set citations were used to create the matrices used as input to the clustering. Next, the figure again reveals the focus of the JKB on agricultural aspects. Additionally, it shows that the all-citing set has a much smoother distribution of papers over the clusters. The coverage of the *Ecology, economics and impact cluster* for example, is at the same level as the publications in *Forest Ecology*, *Agriculture* and (more to the left) *Business Management*, and *Sociological aspects.* This is in line with the bias of the seed set we determined above, when we investigated the focus of the set using fields.

9.3.4 Most cited publications

Finally, since the HCKB is made up of highly cited publications, we take a look

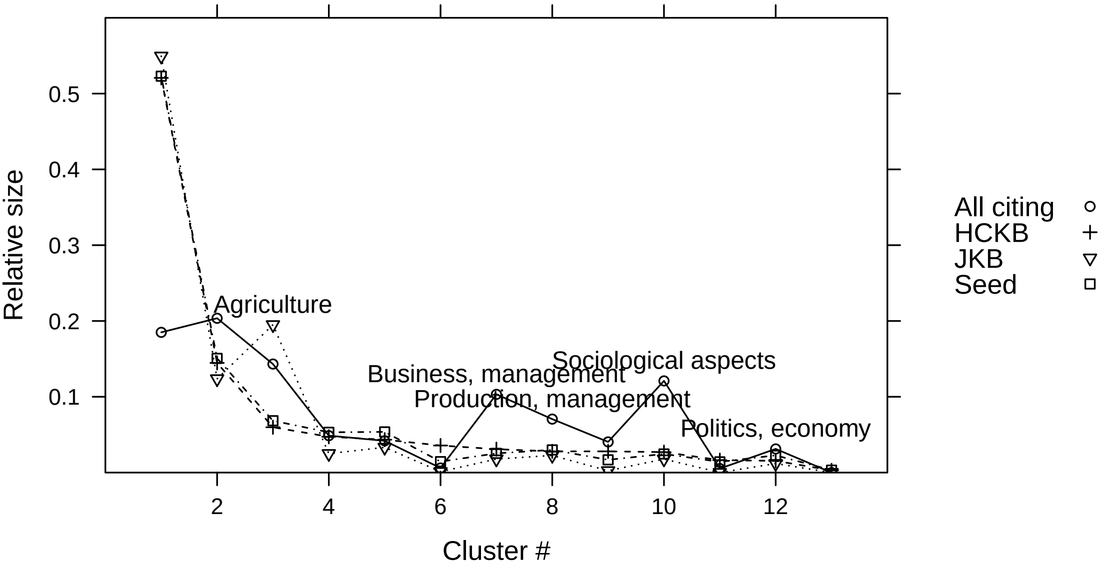


Figure 26 Share of the publications in the clusters from the HCKB, and the number of citing publications from the seed set, the set citing all HCKB publications, and the JKB.

at the best cited publications. First, in **Table 14**, we show the overall best cited publications in the HCKB, as measured by the CPP/FCSm. It shows that these publications are so-called 'classics', as they are relatively old. The first publication is the main reason why *Sociological Aspects* is an outlier in **Figure 24**. Interestingly, these publications are mostly methodological papers and papers related to sociological research (while this was underrepresented in our analyses above) and we find the first environmental publication only at the tenth place (not in the table). The table also shows that the absolute number of citations does not correspond linearly with the CPP/FCSm. For example, the second publication has less citations than the third, yet when compared to its own environment, it is doing better (has a higher CPP/FCSm) than the third publication.

If we restrict the publication year to *after* 1999, we get a different, probably more representative illustration of the important publications. **Table 15** contains the five best cited publications published after 1999. This table shows more publications which appear to be related to what SD is about, although the first two are (again) methodological publications. Moreover, we see that a number of seed publications also appear in this list of highly cited publications. Still there is a contrast in the type of research that is visible in this table. One of the reasons for this contrast could be that the research in sustainability science has evidently roots in sociological research, but that these roots are supporting different types of research at the moment. Also, the coverage of sociological research in the Web of Science database, as well as the

Cluster	Title	Journal	Year	C	CPP/FCSm
Sociological Aspects	The moderator mediator variable distinction in social psychological-research - conceptual, strategic, and statistical considerations	Journal of Personality And Social Psychology	1986	9,350	292
Econometrics	Cointegration and error correction – representation, estimation and testing	Econometrica	1987	3,728	215
Production, Business	The iron cage revisited – institutional isomorphism and collective rationality in organizational fields	American Sociological Review	1983	2,539	191
Business, Management	Economic-action and social-structure – the problem of embeddedness	American Journal of Sociology	1985	2,984	182
Business, Management	Social capital in the creation of human-capital	American Journal of Sociology	1988	2,223	173

Table 14 The title, journal name, and publication year of the five best cited publications in the HCKB together with the cluster in which these publications appear.

importance of books in sociology, may influence the more recent research in the knowledge base.

Cluster	Title	Journal	Year	C	CPP/FCSm
Econometrics	Testing for unit roots in heterogeneous panels	Journal of Econometrics	2003	205	56
Ecology, Economics, and Impact	The third wave of science studies: studies of expertise and experience	Social Studies of Science	2002	84	42
Agriculture	The main elements of sustainable food chain management	Cereal research communications	2006	23	40
Ecology, Economics, and Impact	Is the construction sector sustainable? Definitions and reflections	Building research and information	2006	11	40
Forest ecology	Biodiversity hotspots for conservation priorities	Nature	2000	1648	38

Table 15 The title, journal name, and publication year of the five best cited publications *published after 1999* in the HCKB together with the cluster in which these publications appear.

9.3.5 Diversity in the clusters

Finally, we calculated the diversity of the origin of the publications citing the publications in the clusters. **Figure 27** represents the results. For some of the smaller clusters, the result is rather noisy, with lines that are less smooth. As could be expected, in the more specialised clusters, we see that the diversity is less than in the larger clusters. However, the general trend appears to be an increasing diversity. As a means to verify this, we included the diversity of the seed set as well, but now based on the originating fields of the publications in the seed set instead of the fields of the citing publications. This line can be seen at the top of the graph and it follows the same general trend towards a more diverse attribution to fields.

In our opinion this implies two things. First, the publications in the HCKB are increasingly interesting for publications from different fields. Second, the HCKB is still firmly based in different fields, and a trend to a specialised KB is not yet visible. For an interdisciplinary field such as SD this is not necessarily a negative development, as the scientific recognition of the researchers is still based in their originating field, an observation we also made for other converging fields (Buter et al., 2010a, 2010b). However, if the field of sustainability becomes more and more established we expect the diversity to reduce, particularly in the more specialised clusters.

9.4 Summary of main findings and concluding remarks

We presented an analysis of important aspects of the knowledge base of publications which are enabling sustainability science. Our approach started by locating publications that specifically indicate their relation to sustainability or sustainable development by mentioning either (or both) of these terms in their title. From the resulting ‘seed set’ of publications we then identified the cited publications which make up the knowledge base. Subsequently, we selected the part of the knowledge base that was highly cited, by both the publications from the seed set and the larger research environment (the other citing publications in the Web of Science database). The result was a highly-connected set of papers. By comparing the orientation of this set with other sets, in terms of fields and clusters, we were able to point out differences between the research present in these different sets.

First we investigated the orientation of the HCKB by comparing the coverage of fields and clusters in the HCKB in the seed set, the set of all publications citing the HCKB, and the set of the publications which make up the knowledge base of the specialised sustainability journals. We found that particularly the orientation of the last set was different from the HCKB. Taken as a whole, the JKB shows an orientation on agricultural aspects. Moreover, some fields are under-represented (*Economics, Management*), while others are over-represented (*Energy & Fuels, Agronomy, Soil Science*). Additionally, the coverage of important multidisciplinary journals such as Nature and Science (in this study represented by the field *Multidisciplinary Sciences*) is relatively low in the JKB. As a result, we think that the research published in these specialised journals does not represent the full, multidisciplinary spectrum of sustainability research.

Although orientation and methodology of our paper is different in important aspects, there are some similarities with the approach taken by Kajikawa et al. (2007), and it is interesting to compare our results. First they find a significantly higher number of publications because they also look for ‘sustainab’ words in the abstract

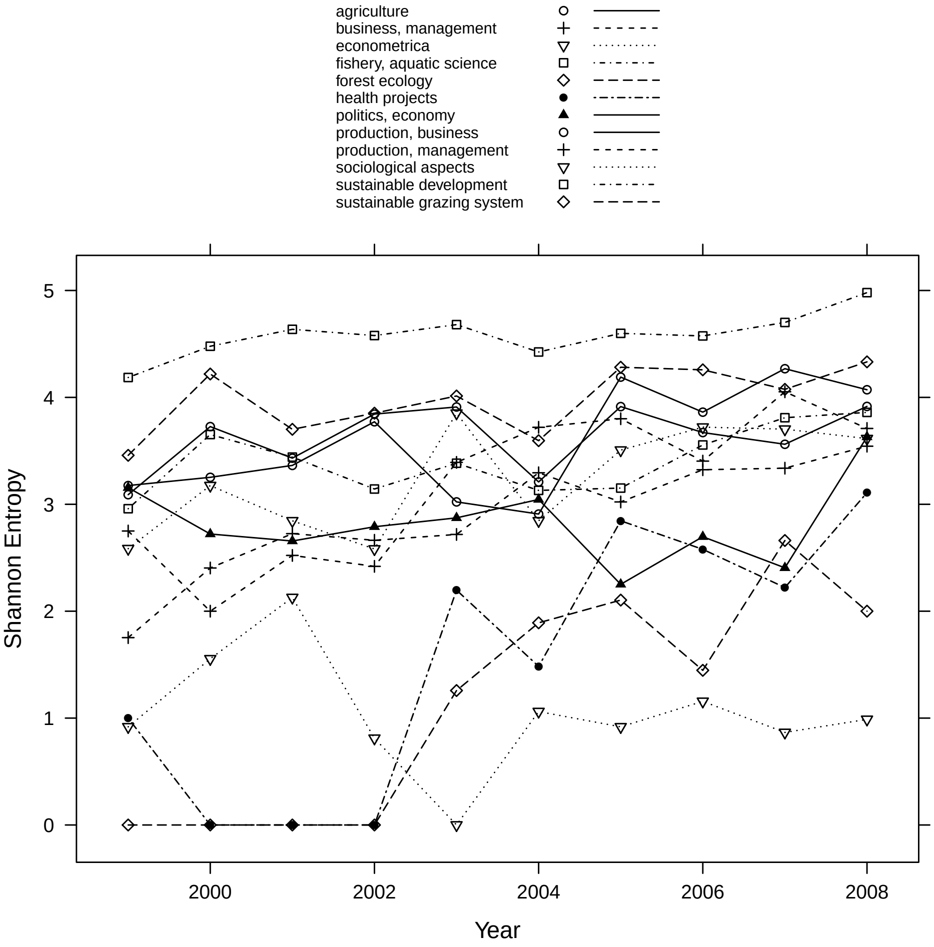


Figure 27 Diversity as measured as the Shannon entropy of the relative occurrence of the fields of citing the publications in the clusters; for the seed set we used the originating fields of the publications themselves.

of publications, whereas we onbly consider the title, as we regard this as an even stronger indication of the intention of the authors. Kajikawa et al. (2007) arrive at about the same number of clusters as in this study, but their clusters are based on a different set of publications with different characteristics. The most striking difference is the distribution of the number of publications in the clusters. We find one large cluster that accounts for over fifty percent of the clustered publications, whereas the largest cluster found by Kajikawa et al. (2007) accounts for only sixteen percent. Nonetheless, many of their clusters also appear in our list, but in less detail. In fact, our result is more similar to the list in a further study of Kajikawa (2008), with an exception for the business and production focus identified in our study which is present in less detail in the Kajikawa list. The diversity of the clusters is measured differently by Kajikawa and colleagues, and they do not show the development over time. We also notice that they present higher journal weight factors for the more specialised clusters, corresponding with the lower diversity found by us in the specialised clusters.

We showed that the rapidly developing field of sustainability science has various interdisciplinary pillars on which its current research is built. With different fields come different subjects, but also different vocabulary (Palmer, 1999), culture (Ruiz-Baños et al., 1999), and even different customs with respect to writing titles (Buter and Van Raan, 2011). This, together with other aspects such as time and money (Kostoff, 2002 and Kajikawa, 2008), make it a challenge to create a new field. Although it is clear that this is required for the field of sustainability science, we currently do not see an increase in the shared use of the HCKB. Additionally, we do not see an increased used of clusters in the four meta-clusters. If the HCKB is a good representation of the knowledge base of the budding field of sustainability science, then we can conclude that the diversity of the citing publications shows that the field is still highly interdisciplinary–and becoming more so on average.

Our current approach focused mostly on the highly cited knowledge base (HCKB) of our seed set. Future research may refine this by using a more elaborate approach for 1) creating the relevant knowledge base; and 2) extracting the relevant citing literature. For example, we can introduce more steps between the initial 'guess' (our seed set) to the final 'seed', but we could also expand the set of citing publications in more than one iteration. Another possible extension of our research is to extract a lexical definition from the seed set, to be used on databases which do not have the citation data (readily) available (e.g., patent databases, INSPEC, PubMed). Additionally, we know that much of SD is captured in grey literature, and a logical extension of our approach would therefore be to see how scientific literature is referenced from these sources, and to compare it with the picture that arises from the present analysis. Moreover, it would be interesting to focus more on the time dimension by inspecting changes in research topics and research networks within the set of papers, perhaps even identify trends and hypes. Finally, we have not focused on actors in our research, yet they are of course the true 'enabling' elements of sustainability science, and therefore they would merit a full analysis of their own.

Acknowledgements

We kindly acknowledge the help of Bert de Wit, MSc. for bringing sustainability and science to our attention and for his suggestions to earlier versions of this paper. In addition, we are grateful for the assessment of the resulting map by Prof. Herman Eijsackers.

References

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Buter, R. K. and Noyons, E. C. M. (2002a). Using bibliometric maps to visualise term distribution in scientific papers. In *Information Visualisation, International Conference on*, pages 697. Los Alamitos, CA, USA.

Buter, R. K. and Noyons, E. C. M. (2005). Using reference structures to evaluate co-word structures: First explorations. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics*, pages 260–264. Stockholm.

Buter, R. K., Noyons, E. C. M. and Van Raan, A. F. J. (2010a). Identification of converging research areas using publication and citation data. *Research Evaluation, 19*(1), 19–27.

Buter, R. K., Noyons, E. C. M. and Van Raan, A. F. J. (2010b). Searching for converging research using field to field citations. *Scientometrics, 86*(2), 325–338.

Buter, R. K. and Van Raan, A. F. J. (2011). Non-alphanumeric characters in titles of scientific publications: an analysis of their occurrence and correlation with citation impact. *Journal of Informetrics, 5*(4), 608–617.

Clark, W. C. (2007). Sustainability science: A room of its own. *Proceedings of the National Academy of Sciences, 104*(6), 1737–1738.

Clark, W. C. and Dickson, N. M. (2003). Sustainability science: The emerging research program. *Proceedings of the National Academy of Sciences of the United States of America, 100*(14), 8059–8061.

Dodds, S. (1997). Towards a 'science of sustainability': Improving the way ecological economics understands human well-being. *Ecological Economics, 23*(2), 95–111.

Eck, N. J. and Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Elliot, J. A. (2006). *An introduction to Sustainable Development.* Abingdon: Routledge.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821.

Holmberg, J. and Sandbrook, R. (1992). Sustainable development: What is to be done?. In *Policies for a Small Planet*, pages 19–38. London: Earthscan.

Kajikawa, Y. (2008). Research core and framework of sustainability science. *Sustainability Science, 3*(2), 215–239.

Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K. and Komiyama, H. (2007). Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science, 2*(2), 221–231.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters, 31*(1), 7–15.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Grübler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., Mooney, H., Moore, B., O'Riordan, T. and Svedin, U. (2001). Sustainability science. *Science, 292*(5517), 641–642.

Kates, R. W., Parris, T. M. and Leiserowitz, A. A. (2005). What is sustainable development? (cover story). *Environment, 47*(3), 8–21.

Keitsch, M. M. (2010). Sustainability and science – challenges for theory and practice. *Sustainable Development, 18*(5), 241–244.

Kirkby, J., O'Keefe, P. and Timberlake, L. (1995). *The Earthscan Reader in Sustainable Development*. London: Earthscan.

Komiyama, H. and Takeuchi, K. (2006). Sustainability science: building a new discipline. *Sustainability Science, 1*(1), 1–6.

Kostoff, R. N. (2002). Overcoming specialization. *BioScience, 52*(10), 937.

Lee, K. N. (1993). *Compass and Gyroscope: Integrating Science And Politics For The Environment*. Washington: Island Press.

Mayer, A. L. (2008). Strengths and weaknesses of common sustainability indices for multidimensional systems. *Environment International, 34*(2), 277–291.

Moed, H., Bruin, R. D. and Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Moed, H. F. (2005a). *Citation Analysis in Research Evaluation*, volume 9 of *Information Science and Knowledge Management*. Dordrecht: Springer.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool*. PhD thesis, Leiden: Leiden University.

Noyons, E. C. M., Moed, H. F. and Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115–131.

Noyons, E. C. M. and van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science, 50*(3), 242–253.

Pearce, D., Barbier, E. and Markandya, A. (1990). *Sustainable Development - Economics and Environment in the Third World*. Aldershot: Edward Elgar.

Peters, H. P. F., Braam, R. R. and Van Raan, A. F. J. (1995). Cognitive resemblance and citations relations in chemical engineering publications. *Journal of the American Society for Information Science, 46*(1), 9–21.

Phillis, Y. A. and Andriantiatsaholiniaina, L. A. (2001). Sustainability: an ill-defined concept and its assessment using fuzzy logic. *Ecological Economics, 37*(3), 435–456.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann.

Rafols, I. and Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Ruiz-Baños, R., Bailón-Moreno, R., Jiménez-Contreras, E. and Courtial, J. (1999). Structure and dynamics of scientific networks. part i: Fundamentals of the quantitative model of translation. *Scientometrics, 44*(2), 217–234.

Sharachchandra, M. L. (1991). Sustainable development: A critical review. *World Development, 19*(6), 607–621.

Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes, 26*(1), 90–102.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.

Waltman, L., van Eck, N. J. and Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635.

Weinberg, A. M. (1972). Science and trans-science. *Minerva, 10*(2), 209–222.

Part III

General Discussion, Conclusions and Future Prospects

10

General discussion, conclusions and future prospects

10.1 General discussion and conclusions

Many activities in science centre around collecting, selecting, aggregating, creating, and *communicating knowledge*: the scientific discourse is an important factor in the spread, development, and archiving of scientific knowledge. To have a worthwhile discourse, publications need to refer to knowledge in other publications, or in other bodies of knowledge. Such references are an important reason why there is much cognitive and social structure in science. In **Section 1.4.8** we specify a structure in an abstract sense as a collection of possibly overlapping groups of related (elements from) publications. In one set of publications, there is no unique structure, because there are different groupings or constellations of elements possible. On the one hand a constellation depends on the algorithm used to create groups, but also depends on the element that is used: a constellation based on words will be different from a constellation based on journals.

10.1.1 Discussion and conclusions based on Part II

Throughout this thesis, *common ground* emerged as a unifying concept that enables us to explain how research represented by different scientific publications is related and results in *structures*. Although common ground is an abstract concept (there can never be something like *the* common ground of a field), it is nevertheless made up out of concrete elements: jargon, basic scientific literature, or important scientific institutions. For our purposes, these elements need to be represented in scientific publications and should be names or other syntactic constituents which refer to these elements. Subsequently, these references can be interpreted as signs and for example be used to connect to the (cognitive) world view of a researcher. In this dissertation, we have investigated the following, regarding the common ground, its elements and structures.

- The use of visualisation to bring together representations of common ground in a single context as a specific structure.
- The use of maps to translate between world view and representations of the common ground.
- Common ground expectations have been shown in the use of non-alphanumeric characters in titles of publications as they structured related publications.
- The different external representations of common ground have been used to find converging research.
- The common ground of a particular converging field (sustainability) has been investigated through the structure of its scientific knowledge base.

Since visualising structures is a powerful tool to analyse science, we started in **Chapter 3** with a description of a design of a mapping tool. The first objective of our implementation was to overcome limitations of our first-generation interactive maps (Noyons, 1999), in particular the inability to locate familiar items on a map. Additionally, by building a comprehensive tool that both displays maps and gives access to underlying (bibliographic) data, we enabled users to perform different analysis tasks with a single tool. Currently, our implementation is still one of very few bibliometric visualisation tools that can be used in both performance analyses and domain analyses, due to its combined information retrieval and charting capabilities.

The mapping tool also serves as the basis for subsequent applications and analyses. The first application is presented in **Chapter 4**, where we add functionality to visualise the distribution of terms as a spread of numbers and colours over a map. A sample analysis confirmed that words that have a particular meaning in the context of a set of publications highlight particular parts of a bibliometric map, while generic terms are spread over the whole map. The extend of the spread over groups can differentiate between specific and generic terms, which may help to select specific terms to include in a delineation, or to in a cognitive (based on words and phrases), quantitative map.

Spread over a map can also be quantified, for instance using the Shannon entropy[80]. Our use of entropy as the basis to quantify spread was inspired by its application in the machine learning algorithm C4.5 by Quinlan (1993). Based on our experience with the research in **Chapter 4** we further explored this idea in Buter and Noyons (2005)[81]. Recently the Shannon entropy has also been used as one of the indicators to measure spread[82] in Rafols and Meyer (2009).

---

[80] The Shannon entropy is defined as $-\sum_i p_i \log p_i$, where p_i is the chance that item i occurs, for which we commonly employ the relative share of item i in some set I. For example, the spread of an NP n over a map with k elements, would be the vector $(n_1, n_2, ..., n_k)$ that can be converted to a vector of relative counts by dividing by the sum of these counts: $p_i = n_i/\bar{n}$, where $\bar{n} = \sum_i^k n_i$.

[81] We did not include this conference publication in this thesis because of its limited size and subsequent lack of substance. However, we did further develop this idea into a tool that we have applied in CWTS projects to help term selection for delineations as well as maps. The use of this tool is also mentioned in our contributions in **Chapter 9 page 170**.

[82] Actually, the authors use the term ‘coherence’, which is probably the opposite of spread (less spread

So, an important assumption in our work is that a set of publications that is large enough and is associated with some structure that appears sensible (for example, to a field expert) can be regarded as a context in the sense that a common ground provides the cognitive context of a publication. Then, because there are different representations of the same context, structures based on for instance journals can help to assess another structure that is based on for example phrases. A demonstration that this assumption appears valid was given in **Chapter 4**, as well as in Buter and Noyons (2005). Based on this, we again extend in **Chapter 5** our mapping tool with the functionality to look up phrases from publications and group these phrases in a qualitative (subjective, personal) map. By employing this tool and by letting users draw their own maps (pencil and paper style) we confirmed our statement in **Section 1.4.8** that there is no unique structure, and (as a result) we cannot speak of a true or valid map of a research area, because of personal preferences in labelling nodes and fine-grained knowledge of areas of personal interest or expertise. From this we conclude that when maps are used as a bibliometric tool, the utility aspect of a map (see **page 50**) should be considered more important than the structural aspect.

An important consequence of the awareness of using (publish, interact, or research in) a (research) context, is a level of expectation that certain words, references, institutions, or even special characters will occur. For example, our analysis in **Chapter 6** makes clear that the use of non-alphanumeric characters such as quotes and colons in a title, can be considered normal for an average publication in the WOS. Moreover, these publications with titles containing non-alphanumeric characters have a statistically significant larger average impact than publications without. We think that these observations fit the hypothesis that successful publications should strike a balance between convention (level of expectation) and surprise, already voiced by De Mey (1982, p.175). As a result, it is very difficult for a non-established researcher to artificially boost the success of a publication by including arbitrary non-alphanumeric characters in a title, or to use humour extensively. For a more established researcher, the unexpectedness of the title format or use of language would be partly compensated by the established prestige, although it would still not add to the impact of the publication. Perhaps this is one aspect of science that differentiates it from (popular) art.

Such expectations and the related cautiousness (or even conservatism) in a certain context, make interdisciplinary developments such as converging science interesting. Although convergence was originally described in a more conceptual fashion, we successfully translated this concept in Chapters **7** and **8** into tangible scientific developments. In these chapters we search for converging research by trying different representations for context (fields, keywords, and addresses) and overlap or connection (references, co-occurrence, co-authorship). Interestingly, the two search strategies which have strong cognitive aspects (fields and keywords) were successful, while the strategy using addresses was not, even though the address patterns were designed to reflect the (content of the) fields in which they were expected to be active. We assume

---

implies more coherence).

that this is the cause of the large variability in the whole set of department names, which is too large to capture in simple, uncomplicated patterns. However, this could also imply that department names do not reflect the research context in which their researchers are active, and that names are mostly associated with a certain context by implication, i.e. because they happen to appear frequently in publications published in that context. Another conclusion is that, although we successfully located developments using JSCs as proxies for fields, it would be interesting and rewarding to create more fine-grained divisions of related research (echoing a conclusion of Rinia (2007)), also because this would allow us to search for new developments *within* JSCs. We also investigated the use of the distribution of the growth indicator to address the problem of locating fast growth. The problem of 'large enough' developments is however more difficult to address objectively and remains an interesting problem for future research. Finally, based on both studies into detecting converging science, we also conclude that we were not able to find successful yet unexpected combinations of research areas, like the examples presented in Roco and Bainbridge (2003). Although we cannot exclude that this is due to our particular methodology, we think it more likely that this can be attributed to the preference for explainable (and usually small) steps in scientific developments. Again, these conclusions are in line with what Rinia (2007) concluded after a more extensive study into interdisciplinarity.

Interdisciplinary developments require a strong incentive to overcome challenges that exist when researchers venture into a new research context. Societal push is an example of such an incentive, and sustainability science is an area that is the result of a societal push for sustainable development. Our research in **Chapter 9** was able to show that the three 'pillars' of this area are indeed economics, social sciences, and environmental science. However, a more interesting finding was that the groups in the structure appeared (still) no signs of increased integration. Although this finding requires more thorough investigation and validation, it could imply that external (to science) incentives for integration have more trouble in achieving integration than internal incentives[83]. We also found that for this type of research, exemplars are available at different levels of aggregation or abstraction: JSCs are useful to get a general overview of the contents of a set of publications, while a map gives more detail. Highly cited references also appear to reflect a higher level of abstraction, especially older references. Although more recent references appear more concrete, we conclude that highly cited references are less useful as exemplars for specific research topics: they are a representation of a more general research context. Still, this also needs to be confirmed for other research areas.

10.1.2 Discussion of current use of the mapping tool

We already noted that most of our research has been the result of projects that

---

[83] This conclusion of course also echoed in the remark from Persson (1999) on buildings and bricks, which we quoted in **Section 1.4.6** at **page 38**.

served actual policy demand. The mapping tool has been used in a large number of these projects, most prominently as part of a web application used to interactively go through the detailed numerical results of bibliometric performance analyses[84]. The added value of a map in these analyses is twofold: 1) to visualise a structure in the set of publications being analysed; and 2) to provide a more detailed context for an indicator. In **Figure 28** we show a screen shot of the application, in which the activity of Denmark is visualised in (a map of) neuroscience in a similar fashion as what we describe in **Chapter 4**. The result is more informative than a single indicator for a set of publications. Over the years, institutions such as the Danish Lundbeck Foundation; the EU; the Dutch Council for Spatial, Environmental, and Nature research (RMNO); and Delft University of Technology; have made use of these maps to assess research performance.

From the experience in guiding clients in the use of these maps we can conclude two things. On the one hand it confirms the utility of maps as a general-purpose tool for domain analysis, and for adding context to indicators. On the other hand however, we also learnt that to let a user make full use of the visualisation and the mapping tool, this user needs to confirm to at least two of following three requirements:

- extensive training in the use of the tool;
- interest in, and knowledge of, the research being visualised;
- an interest in using the tool for a clear and well defined purpose.

Usually, this is difficult to achieve, and the most common use of our tool was the display of a map as circles, with the ability to colour these circles to indicate relative output, number of citations and impact, like the example we show in **Figure 28**.

10.2 Future research directions

We consider two important directions of future research. One is the improvement of our visualisation and mapping tool. The other is the expansion of our data sources beyond the bibliographic databases on which we have based most of our current research.

10.2.1 Visualisation and mapping tool

Despite the lack of enthusiasm in the conclusion of the previous section, we are still convinced that visualisation is an important tool for the analysis of complex domains such as science. Not only does a map give users the ability to scan unknown terrain easier, but one does not need to be a field expert to *use* a map. Moreover, recently Vibert et al. (2009) showed that a lack of background knowledge of a particular

---

[84] Most of these projects are collected at `http://studies.cwts.nl/projects`

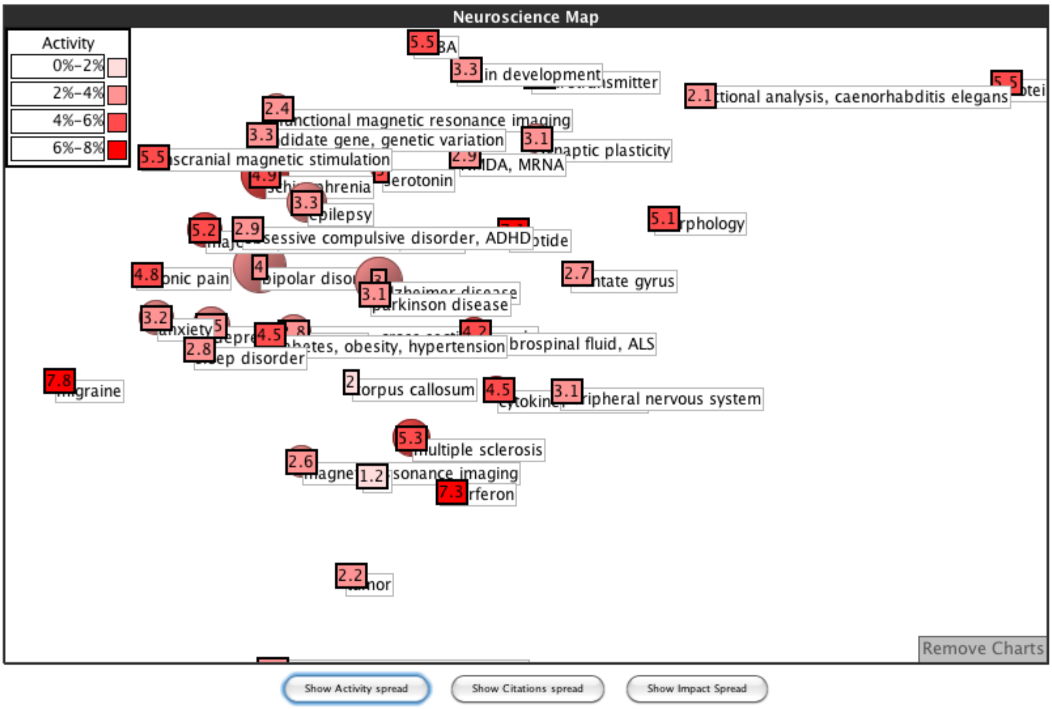


Figure 28 A map of neuroscience that is used to give a detailed overview of the activity of Denmark in neuroscience topics. The activity is measured as a percentage of the total number of publications in the delineation (assumed to represent the world output), and is plotted as a number on top of a cluster. The map is built using NPs. The area of a circle represents the total world output associated with a cluster in the last two years of analysis. The colour of a circle also reflects the relative output: as the legend in the top-left of the figure indicates, a high relative output (6–8 percent) is associated with a dark red colour, while a small percentage (0–2 percent) is light red.

research area does not necessarily hamper the ability to find relevant publications, only the efficiency in which one is able to do so is affected; in our experience, this result is also applicable to the use of maps and other visualisations. However, the complexity of a subject like science also demands more of a mapping tool than we currently provide: a tool should be more flexible while at the same time give users a better sense of where they are looking at.

With respect to *flexibility*, we think that the tool should be more like a telescope then is now the case: a telescope is not only able to show structures at increasing depths, but *also* for different parts of the frequency spectrum (e.g. visual light, ultraviolet, or infrared). In fact, we think that the idea should be abandoned that a single map is used to analyse a science domain. Even more, we think that a user should be able to *influence* which publications are part of the visualisation, by disregarding or adding publications interactively. For this to be useful, once a set of publications is modified, it should (if required) be clustered and laid out again, preferably impromptu. Also, a user should be able to zoom in onto a particular area of the visualisation, or restrict the visualisation to a particular group of publications that appear as a single element in a map. These are two separate actions: the latter is altering the set that is being visualised to only those in the element to be ‘opened’, again requiring impromptu clustering and layout, while the first is only changing the viewport of the visualisation. However, for both actions, a user should have the ability to quickly restore the previous view, i.e. ‘zoom out’ again. Of course, it should also be possible to save a snapshot of a current views. The ability to influence a set of publications being visualised also mitigates some challenges in delineating a field, as it allows users to fine-tune using their expert knowledge or personal preferences or interests. Next to this, a user should be able to change the *type* of element on which a visualisation is based, for instance from co-cited journals to co-occurring phrases, akin to changing the frequency spectrum that is captured by a telescope. To accomplish this, additional research into smart selection algorithms for relevant elements is required, especially for the quick selection of phrases.

Next to the increased flexibility, a future application should have a more central role for the *context* of a set of publications that is being visualised. One way is to use bread crumbs (e.g. De Heer and Card (2004)) to visualise a search path or the ‘zoom context’. Another way is by showing important exemplars for a set of publications that is currently shown. Next to this automatically generated context, a user should be allowed to manually annotate a visualisation and save these annotations for future use, as well as be able to share them with other users, in combination with the saved visualisation snapshots.

Most of the basic tools and techniques to do this are currently available, but in the form of scientific publications, stand-alone (mapping) implementations, or parts of (generic) statistical applications. To assemble the relevant parts, research should probably expand into fields such as information visualisation, and the design of user interfaces, but also into statistics and computer science, for example to find efficient algorithms to select the set most relevant elements to display or to quickly and impromptu cluster that set. These parts will have to be assembled into a single application, like the current tool. However, we envision an opinionated piece of software (Fried et al., 2009, chapter 4) in which we embed best practises to bring a user from goal to result, instead of a general toolbox of independent or loosely coupled tools, such as currently provided in our mapping tool (and akin to graphical tools like Photoshop or Picasa). The best practises codified in the tool should represent our experience in using the mapping tool in projects and by conducting the research described in this thesis.

10.2.2 Expansion of data sources

Future research should also expand beyond the general bibliographic databases such as WOS and Scopus. For example, we are currently investigating the scientific basis of the IPCC[85] Fourth Assessment Report (AR4). Not only does such an investigation serve a concrete societal interest (given the current debate on climate research), but the report also contains many references to grey literature and could thus be an excellent starting point for this line of research. Added to that, we can also start to investigate how we can include relevant data from outside bibliographic databases, for example by including information from open access tools such as Scholar from Google. Of course, including new sources for scientific data introduces challenges with respect to the quality of the acquired data, as for some added data sources, we can no longer rely on peer review to ascertain a basic level of quality, nor will we always have references between publications to serve as a proxy for quality. One way of addressing this is by ensuring that an authoritative source exists for a particular publication, such as the AR4 report, before including it in our analysis. Of course, there is no guarantee that such authoritative source exists, or that it does not also refer to individual publications of inferior quality. This is similar to many peer-reviewed journals, where there is also no guarantee that any individual contribution is necessarily of higher quality than rejected publications, although we can probably assume so in the general case.

Web pages can also serve as an additional source of information, although the problem of authority exists here as well (perhaps even more so). When we include web pages though, we should move beyond the straightforward counting of the number of times a specific word appears on web pages, or the number of hops between one page with a particular word or author name, and another. Instead, contextual information from pages should be automatically combined with scientific publications; this could be words or phrases from pages, references to publications, or author names. We could perhaps even use a summary of a page for this purpose, but this would require us to expand our research into ways of aggregating content. More general research into full-text processing of scientific publications could accompany this research line as well. We could also built on the work we have done on the combination of qualitative and quantitative maps, but now also as a way to organise content from web pages, or reports, next to scientific publications.

References

Buter, R. K. and Noyons, E. C. M. (2005). Using reference structures to evaluate co-word structures: First explorations. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics*, pages 260–264. Stockholm.

---

[85] The IPCC is the *Intergovernmental Panel on Climate Change*, which assesses the scientific, technical and socio-economic information relevant for the understanding of the risk of human-induced climate change (quoted from `www.ipcc.ch/organization/organization.shtml`).

De Heer, J. and Card, S. K. (2004). DOITrees revisited: scalable, space-constrained visualization of hierarchical data. In *Proceedings of the working conference on Advanced visual interfaces* AVI '04, pages 421–424. New York, NY, USA.

De Mey, M. (1982). *The Cognitive Paradigm.* London: D. Reidel Publishing Company.

Fried, J., Heinemeier Hansson, D. and Linderman, M. (2009). *Getting Real: the Smarter, Faster, Easier Way to Build a Successful Web Application.* 37signals.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool.* PhD thesis, Leiden: Leiden University.

Persson, G. (1999). Mistra's role and some lessons learned in the first five years. *Policy Sciences, 32*(4), 323–326.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* Morgan Kaufmann.

Rafols, I. and Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rinia, E. J. (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer.* PhD thesis, Leiden: Leiden University.

Roco, M. C. and Bainbridge, W. S., editors (2003). *Converging Technologies for Improving Human Performance.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Vibert, N., Ros, C., Bigot, L. L., Ramond, M., Gatefin, J. and Rouet, J. (2009). Effects of domain knowledge on reference search with the PubMed database: An experimental study. *Journal of the American Society for Information Science and Technology, 60*(7), 1423–1447.

Summary of this thesis

In this thesis we are concerned with structures in science, as part of a quantitative analysis of scientific publications extracted from the Web of Science database. We abstractly define a structure as a collection of possibly overlapping groups of (elements from) publications. Examples include structures based on co-occurring words and phrases, co-cited journals, or bibliographically coupled authors. An important reason for the existence of these structures is the scientific discourse, the central process in science that is important for spreading, developing and archiving knowledge: to have a worthwhile discourse, it is required to be able to indicate the context of the discourse, by referring to external knowledge, either explicitly available in other publications or implicitly in the common ground of a research area or field. As such, scientific publications provide an excellent basis for the analysis of such structures.

The thesis is divided in three parts. In Part I we give a general introduction to the research background and a bridge that links this background to the core publications in Part II. The publications in this second part can be divided in two groups: the first group of four publications focuses on the use and visualisations of structures, and the second group on finding and analysing converging science. In Part III we describe some general conclusions, discuss our results and conclusions, give some directions for future research, as well as provide the summary you are reading now.

The power of visualisations such as maps is grounded in the use of the powerful processing capabilities of people to quickly correlate visualised information with familiar, internal concepts and use the visualised relations to corroborate existing links or create new links between concepts. The basic assumption on which this use is based, is that an element in a map, corresponding to a group of items in a structure, represents an aspect of a research context, such as a topic or a group of collaborating actors. Our first paper in **Chapter 3** is on the design of a mapping tool that addresses some limitations of earlier generation mapping tools. Our design is of a comprehensive tool that enables a user not only to visualise a map, but also to show charts on selected parts a map, and to retrieve information about the publications represented by elements in the map. The next two papers describe two applications and extensions of this tool. The first application in **Chapter 4** shows that a structure displayed by a map can visualise the difference between generic and specific terms, by colouring map elements with a colour that corresponds to the frequency a term in (publications associated with) an element. An application of this is to select specific terms as part of a delineation or a cognitive map (based on terms). The second application of the mapping tool in **Chapter 5** combines a quantitative (bibliometric) map with a qualitative (not automatically generated) map. This qualitative map is used to collect and order alternative labels for concepts related to those preferred by a user, created by searching for publications with preferred labels and extracting possible alternatives from titles and abstracts. We can view this application as a way to create a translation between intangible concepts and tangible representations of these concepts (or closely related ones) in (titles and abstracts of) publications. The paper also presents two studies into the use of qualitative maps, which show that users prefer network-like maps, and that for one particular field, and one particular type of item (words and phrases), users create different maps with different preferences for topics and detail. From this we conclude that judgements regarding the utility of a map should be preferred over judgements on the validity of a map as a true or valid representation of a research context. Our last paper in this group is **Chapter 6**, in which we investigate the occurrence of non-alphanumeric characters in titles of publications and the correlation of this use with impact, both at the general level of science and at the level of science fields. We show that the occurrence of these non-alphanumeric characters is comparable over related fields and that some fields have started to attribute specific meaning to some of these characters. Additionally, we show that the correlation with impact is positive in general, but only when the most frequent non-alphanumeric characters are included, implying that these frequently occurring characters should be considered normal. We also show that this correlation between the occurrence of non-alphanumeric characters and impact can be absent or even reversed in specific science fields. In general, we conclude that operating within a specific research context comes with expectations regarding content and format, as well as a (general) level of cautiousness regarding diversions from such expectations.

The second group of publications focuses on finding and analysing converging science, which are developments that bring together parts of science that were previously not (or hardly) connected. Such interdisciplinary developments are interesting because they have a general appeal to policy makers, but also because they represent successful translations of knowledge or methodology into a new research context. Such a translation is not straightforward, because it requires researchers to know and successfully address the cautiousness and expectations in a new research context. The first publication in **Chapter 7** describes three search strategies to find converging research. The strategy that uses fields to represent context, as well as the one that uses content (keywords) were both successfully applied to find converging research that was assessed and accepted by field experts. However, a third strategy based on address patterns did not yield useful results. One reason for this might be that the global diversity in address (department) names is too large to capture in simple patterns, but it could also be that a name is only loosely coupled with the research in a particular field. The next publication in **Chapter 8** describes how we improve the search strategy based on fields, by developing a growth indicator that is based on a relative number of citations between fields instead of an absolute number. Also, we try to find a more objective approach to differentiate between normal and fast growth, by using the distribution of the growth indicator. Next to that, we employ a visualisation of growth, and look for indicators for coherent developments, like spread over journals. Again, we successfully find converging research, but also developments that are unidirectional instead of bidirectional. In **Chapter 9**, the last chapter of Part II, we investigate the knowledge base of 'sustainability science', a converging field that is the result of a societal interest in sustainable development. We show that this knowledge base can be divided in three parts: in one part research from economics is dominant, in another research from the social sciences, while in the third research from the environmental studies. A diversion along these lines ('profit, people, and planet') is a well-known characteristic of this field which we successfully reconstruct from the knowledge base. However, the elements of this structure do not appear to show increasing coherence, a trend that we would expect in a converging field, and which may indicate that the 'pillars' still operate as separate fields, and that a transdisciplinary approach to sustainable development is not (yet) attained.

The main conclusion of this thesis is that the structure as we have used it in this thesis, is a versatile abstraction in bibliometric analyses. The elements within a cognitive structure can be regarded as a tangible representation of knowledge, and can be used to filter words or look up alternative labels for a concept. The groups within a structure can be regarded as research context, with expectations about the content and format of research. At a much higher level of aggregation, fields can be regarded as (macro-) structures as well, with also at this level expectations regarding content and format, even in the use of non-alphanumerical characters in titles. Combining research from different contexts is a fruitful basis for new developments in science. And although for that reason the combination of existing structures appears a good idea from a policy point of view, this does not imply that such structures quickly merge into a single, new structure.

Samenvatting - Dutch Summary

In deze dissertatie staan structuren in de wetenschap centraal, als onderdeel van de kwantitatieve analyse van wetenschappelijke publicaties uit de Web of Science database. We definiëren structuren op een abstracte manier als een verzameling van mogelijk overlappende groepen van (elementen uit) publicaties. Voorbeelden zijn structuren die gebaseerd zijn op woorden of frases die samen voorkomen, journals die samen geciteerd worden of auteurs die bibliografisch gekoppeld zijn. Een belangrijke reden voor het bestaan van deze structuren is de wetenschappelijke discussie, het centrale proces in de wetenschap dat belangrijk is voor het verspreiden, ontwikkelen en archiveren van kennis: om een zinnige discussie te voeren is het noodzakelijk om de context van ervan te kunnen aangeven door te verwijzen naar externe kennis, hetzij expliciet aanwezig in andere publicaties, dan wel impliciet in de *common ground* van een onderzoeksgebied of wetenschapsveld. Wetenschappelijke publicaties zijn dan ook een bijzonder goede basis voor de analyse van dit soort structuren.

Het proefschrift is verdeeld in drie delen. In Part I beginnen we met een algemene beschrijving van de wetenschappelijke achtergrond van ons onderzoek, waarna we in een volgend hoofdstuk deze achtergrond verbinden met de publicaties in Part II. De publicaties in dit tweede deel kunnen in twee groepen worden verdeeld: de eerste groep van vier publicaties richt zich op het gebruik en de visualisatie van structuren en de tweede groep op het vinden en analyseren van convergerende wetenschap. In Part III beschrijven we algemene conclusies, bespreken we onze resultaten en conclusies, beschrijven we mogelijkheden voor toekomstig onderzoek en geven de samenvatting die u nu leest, voorafgegaan door de Engelse versie.

De kracht van visualisaties zoals kaarten is de combinatie van het visuele verwerkingsvermogen van mensen om snel patronen te kunnen herkennen en deze te verbinden met bekende concepten, om zo bestaande verbindingen te bevestigen of nieuwe relaties te leggen. Hierbij nemen we aan dat een element op een kaart, die een groep van items in een structuur representeert, een (deel van) een onderzoeksomgeving voorstelt, zoals bijvoorbeeld een onderwerp of een groep samenwerkende actors. Onze eerste publicatie in **hoofdstuk 3** beschrijft het ontwerp van een programma voor wetenschappelijke kaarten (een *mapping tool*) dat een aantal beperkingen aanpakt van eerdere generaties van dit soort software. We hebben een ontwerp gemaakt dat een gebruiker niet alleen in staat moet stellen om een kaart te visualiseren, maar ook om grafieken op elementen in de kaart te kunnen tonen en om informatie over de onderliggende publicaties op te kunnen halen. De volgende twee publicaties beschrijven uitbreidingen en toepassingen van dit programma. De eeste toepassing in **hoofdstuk 4**, toont dat een structuur die wordt afgebeeld in een kaart het verschil tussen generieke (algemene) en specifieke termen kan visualiseren, door een element op de kaart een kleur te geven die overeenkomt met het aantal keren dat een term voorkomt in (publicaties) binnen dit element. Een toepassing hiervan is het selecteren van specifieke termen om deel uit te maken van een afbakening of van een cognitieve kaart (die groepen van termen toont). In de tweede toepassing van de *mapping tool* in **hoofdstuk 5**, wordt een kwantitatieve (bibliometrische) kaart met een kwalitatieve (niet automatisch gegenereerde) gecombineerd. Deze kwalitatieve kaart wordt gebruikt om alternatieve labels voor concepten te verzamelen en te ordenen die te maken hebben met labels die de voorkeur hebben van een gebruiker, door publicaties te zoeken met de voorkeurslabels en zo mogelijke alternatieven in de titels en samenvattingen kunnen halen. Deze toepassing kunnen we zien als een manier om een vertaling te maken tussen niet-tastbare concepten en tastbare representaties van deze concepten (of gerelateerde varianten) in (titels en samenvattingen van) publicaties. Het hoofdstuk bevat ook nog twee studies naar het gebruik van kwalitatieve kaarten, waaruit blijkt dat gebruikers een voorkeur hebben voor netwerkachtige kaarten en ook dat voor eenzelfde veld en element (woorden of frases), gebruikers totaal verschillende kaarten maken, met verschillende voorkeuren voor labels en detail. Hieruit concluderen wij dat uitspraken over de bruikbaarheid van een kaart de voorkeur moeten hebben boven uitspraken of een kaart een getrouw beeld geeft van een onderzoeksomgeving. Onze laatste publicatie in deze groep is **hoofdstuk 6**, waar we het gebruik van niet-alfanumerieke tekens in titels van publicaties onderzoeken en de correlatie hiervan met impact, zowel in het algemeen, als voor specifieke onderzoeksvelden. We tonen aan dat het gebruik van deze niet-alfanumerieke tekens vergelijkbaar is voor velden die met elkaar te maken hebben en dat auteurs in sommige velden een specifieke betekenis zijn gaan geven aan sommige van deze tekens. Daarnaast laten we zien dat er in het algemeen een positieve correlatie bestaat met impact, maar alleen wanneer de meest voorkomende niet-alfanumerieke tekens worden meegenomen in de analyse, wat betekent dat deze vaak voorkomende tekens als normaal moeten worden beschouwd. We laten ook zien dat deze correlatie afwezig of zelfs omgekeerd kan zijn in bepaalde wetenschapsvelden. In het algemeen concluderen we dat het werken in een bepaalde onderzoeksomgeving samengaat met een zekere verwachting aangaande inhoud en vorm, naast een (algemene) behoedzaamheid af te wijken van dergelijke verwachtingen.

De tweede groep publicaties richt zich op het vinden en analyseren van convergerende wetenschap, wat ontwikkelingen zijn die delen van de wetenschap bij elkaar brengen die daarvoor niet (of nauwelijks) met elkaar in verbinding stonden. Zulke interdisciplinaire ontwikkelingen zijn interessant, omdat ze een zekere aantrekkingskracht hebben voor beleidsmakers, maar ook omdat ze succesvolle vertalingen representeren van kennis of methoden naar een nieuwe onderzoeksomgeving. Een dergelijke vertaling is niet eenvoudig, omdat onderzoekers hiervoor de behoedzaamheid en verwachtingen in een nieuwe onderzoeksomgeving moeten onderkennen en verwerken. De eeste publicatie in **hoofdstuk 7** beschrijft drie zoekstrategiën om convergerend onderzoek te vinden. De strategieën die gebruik maken van wetenschapsvelden en van inhoud (*keywords*) leveren allebei resultaten op die door experts zijn getoetst en geaccepteerd. Echter, een derde strategie gebaseerd op adrespatronen levert geen bruikbare resultaten op. Een reden hiervoor kan de wereldwijde diversiteit in (faculteits-) namen die te groot is om in eenvoudige patronen te vatten, maar het kan ook zijn dat de naam van een faculteit slechts beperkt verbonden is met het onderzoek in een bepaald veld.

De volgende publicatie in **hoofdstuk 8** beschrijft hoe wij de op velden gebaseerde zoekstrategie verbeteren, door een groeiïndicator te ontwikkelen die is gebaseerd op een relatief aantal citaties in plaats van een absoluut aantal. Ook proberen we een meer objectieve manier te vinden om onderscheid te maken tussen normale en snelle groei, door gebruik te maken van de verdeling van de groeiïndicator. Daarnaast gebruiken we een visualisatie van de groei en zoeken we naar indicatoren voor coherente ontwikkeling, zoals verdeling over journals. Ook nu vinden we weer convergerende wetenschap, maar ook ontwikkelingen die slects een kant op werken, in plaats van twee kanten. In **hoofdstuk 9**, de laatste publicatie in Part II, onderzoeken we de *knowledge base* van *sustainability science* (duurzaamheidswetenschap), een convergerend veld dat het gevolg is van het belang dat de samenleving hecht aan duurzame ontwikkeling (*sustainable development*). We tonen aan dat deze *knowledge base* in drieën kan worden verdeeld: in een deel is economisch onderzoek met name aanwezig, in een ander sociologisch onderzoek, en in een derde onderzoek naar landbouw, visserij en milieu. Een dergelijke driedeling is een bekende karakteristiek van dit onderzoek en staat ook bekend als *profit, people, en planet*, welke we dus kunnen reconstrueren uit de *knowledge base*. Opvallend is dat de groepen uit deze structuur geen toenemende integratie vertonen, iets wat we zouden verwachten in een convergerend veld en wat er op kan duiden dat de drie „pilaren” nog steeds zich als onafhankelijke velden gedragen en dat een transdisciplinaire benadering van duurzame ontwikkeling (nog) niet is bereikt.

De belangrijke conclusie van deze dissertatie is dat de *structuur* een veelzijdige abstractie is voor bibliometrische analyses. Elementen uit cognitieve structuren kunnen bijvoorbeeld worden beschouwd als tastbare representaties van kennis, die gebruikt kunnen worden om woorden te filteren of om alternatieve labels voor concepten te vinden. De groepen van publicaties die met deze elementen zijn verbonden kunnen worden beschouwd als onderzoeksomgevingen, met bepaalde verwachtingen over de inhoud en de vorm van onderzoek. Op een veel hoger aggregatieniveau kunnen velden ook worden beschouwd als structuren en ook op dit niveau bestaan er verwachtingen over inhoud en vorm van onderzoek, wat we zelfs kunnen afleiden uit het gebruik van niet-alfanumerieke tekens in titels. Het combineren van onderzoek uit verschillende omgevingen is een vruchtbare bodem voor nieuwe ontwikkelingen in de wetenschap en hoewel de combinatie van bestaande structuren vanuit een beleidsperspectief om die reden een goed idee lijkt, hoeft dit nog niet als gevolg te hebben dat deze structuren ook snel opgaan in een nieuwe structuur.

Postscript

As this thesis slowly took shape, the Internet developed from a novelty into a paradigm shift. Among many things, it allowed for new and improved ways of getting scientific results to the general public. Preprint archives have found their place next to the sites of more traditional publishers and it is now easier than ever before to find scientific publications on specific topics, either through specialised (sometimes closed) portals, or through general search engines. However, there are also less positive effects of the unmoderated character of internet publications. For example, prolific bloggers without any formal scientific training can be quoted alongside publications of respected scientists, as if they were of equal quality. The same goes for news outlets, as both the respected and the obscure can spread opposite claims about scientific topics such as climate change, evolution, or genetics. Without the proper context, many hapless netizens are unable to judge which source (if any) to trust. Still, science has always had to find a balance between political, societal, economic, and creative forces which try to influence its course. Therefore, it has seldom been without controversy, yet has proven to be an immensely powerful, structuring element in our world. This is mainly due to the universality of its basic methodology, its focus on quality, and its strive for continuous improvement. For some, this makes science suspicious and perhaps even something to fear—for most of us though, it will remain a source of opportunities, continuous development, and inspiration.

> *A society made up of individuals who were capable of original thought would probably be unendurable. The pressure of ideas would simply drive it frantic.*
>
> — H.L. Mencken, ‘Minority Report’, 1956, 2006 reprint, p.10

Curriculum Vitæ

Reindert Klaas Buter (roepnaam *Renald*) werd geboren in Hasselt (Overijssel) op 20 april 1973. Na een lagere-schooltijd in Hasselt ging hij naar het Carolus Clusius College te Zwolle, alwaar hij na zes jaar zijn diploma Atheneum-B behaalde. Om van zijn hobby zijn werk te kunnen maken, besloot hij Technische Informatica te gaan studeren aan de Universiteit Twente te Enschede. Tijdens zijn studie raakte hij geboeid door kunstmatige intelligentie en natuurlijke-taalverwerking. Praktijkervaring hiermee kon hij opdoen aan het einde van zijn studie bij Schuitema BV waar hij neurale netwerken gebruikte om verkoopprognoses te maken en bij de Belastingdienst waar hij het fenomeen datamining onderzocht. Na zijn studie ging hij direct aan de slag bij het Centrum voor Wetenschaps- en Technologiestudies (CWTS). Daar werkte hij een lange tijd en hield zich onder meer bezig met het programmeren van een map-interface, het opzetten en onderhouden van *data warehouses* voor bibliometrische data, het vervaardigen van websites en webinterfaces om interactief bibliometrische data te presenteren, maar ook met het ontwikkelen van een plannings- en tijdsbestedingssysteem om de vele projecten van het CWTS te helpen monitoren. Daarnaast hield hij zich bezig met de uitvoering van bibliometrische projecten voor nationale en internationale opdrachtgevers. Ondertussen begon zijn promotieonderzoek onder begeleiding van prof. dr. A.F.J. Van Raan, dat tussen de ontwikkeling van bibliometrisch instrumenten en het uitvoeren van analyses laveerde, langzaam vorm te krijgen, mede door een aantal door de EU gefinancierde en in opzet ambitieuze projecten. In 2010 besloot hij het CWTS te verlaten, maar daar wel zijn promotieonderzoek voort te zetten. Momenteel is hij werkzaam als IT-er en heeft hij ook een eigen bedrijf genaamd *Data Ludentes* (spelende data) waarmee hij zijn jarenlange ervaring op het gebied van *science* en *business analytics* en *intelligence* probeert uit te venten. Hij is sinds 1999 getrouwd met Mirjam en samen hebben zij twee dochters, Yaël en Amarinthe.

Complete list of references

Aho, A. V. (1990). Algorithms for finding patterns in strings. In *Handbook of theoretical computer science (vol. A): algorithms and complexity*, pages 255–300. MIT Press, Cambridge, MA, USA.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*(4), 249–277.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *American Economic Review, 84*(2), 406–411.

Ball, R. (2009). Scholarly communication in transition: The use of question marks in the titles of scientific articles in medicine, life sciences and physics 1966–2005. *Scientometrics, 79*(3), 667–679.

Barber, J. A. and Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine, 19*(23), 3219–3236.

BBC and Horizon (2010). Science under attack. . retrieved 2011-02-21.

Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17. Singapore.

Berg, J. and Wagner-Döbler, R. (1996). A multidimensional analysis of scientific dynamics. *Scientometrics, 35*(3), 321–346.

Besselink, M. G. H., Van Santvoort, H. C., Buskens, E., Boermeester, M. A., Van Goor, H., Timmerman, H. M., Nieuwenhuijs, V. B., Bollen, T. L., Van Ramshorst, B., Witteman, B. J. M., Rosman, C., Ploeg, R. J., Brink, M. A., Schaapherder, A. F. M., Dejong, C. H. C., Wahab, P. J., Van Laarhoven, C. J. H. M., Van Der Harst, E., Van Eijck, C. H. J., Cuesta, M. A., Akkermans, L. M. A. and Gooszen, H. G. (2008). Probiotic prophylaxis in predicted severe acute pancreatitis: a randomised, double-blind, placebo-controlled trial. *The Lancet, 371*(9613), 651–659.

Börner, K., Chen, C. and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255.

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Braam, R. R., Moed, H. F. and Van Raan, A. F. J. (1991b). Mapping of Science by Combined Co-Citation and Word Analysis II: Dynamic Aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Brooks, M. and Campbell, J. (1999). Interactive graphical queries for bibliographic search. *Journal of the American Society for Information Science, 50*(9), 814–825.

Buckland, M. K. (1991). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351–360.

Budd, J. and Hurt, C. D. (1991). Superstring theory: Information transfer in an emerging field. *Scientometrics, 21*(1), 87–98.

Buter, R. K. and Noyons, E. C. M. (2001). Improving the functionality of interactive bibliometric science maps. *Scientometrics, 51*(1), 55–68.

Buter, R. K. and Noyons, E. C. M. (2002a). Using bibliometric maps to visualise term distribution in scientific papers. In *Information Visualisation, International Conference on*, pages 697. Los Alamitos, CA, USA.

Buter, R. K. and Noyons, E. C. M. (2005). Using reference structures to evaluate co-word structures: First explorations. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics*, pages 260–264. Stockholm.

Buter, R. K., Noyons, E. C. M., Van Mackelenbergh, M. and Laine, T. (2006). Combining concept maps and bibliometric maps: First explorations. *Scientometrics, 66*(2), 377–387.

Buter, R. K., Noyons, E. C. M. and Van Raan, A. F. J. (2010a). Identification of converging research areas using publication and citation data. *Research Evaluation, 19*(1), 19–27.

Buter, R. K., Noyons, E. C. M. and Van Raan, A. F. J. (2010b). Searching for converging research using field to field citations. *Scientometrics, 86*(2), 325–338.

Buter, R. K. and Van Raan, A. F. J. (2011). Non-alphanumeric characters in titles of scientific publications: an analysis of their occurrence and correlation with citation impact. *Journal of Informetrics, 5*(4), 608–617.

Calero, C., Buter, R. K., Valdés, C. C. and Noyons, E. C. M. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics, 66*(2), 365–276.

Callon, M., Law, J. and Rip, A., editors (1986). *Mapping the Dynamics of Science and Technology*. The MacMillan Press Ltd, London.

Carayol, N. and Thi, T. U. N. (2005). Why do academic scientists engage in interdisciplinary research?. *Research Evaluation, 14*(1), 70.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141–1164.

Cassirer, E. (1942). The influence of language upon the development of scientific thought. *The Journal of Philosophy, 39*(12), 309–327.

Chalmers, A. F. (1999). *What Is This Thing Called Science?*. Buckingham: Open University Press.

Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine, 15*(13), 1447–1458.

Chavalarias, D. and Cointet, J. (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. *Scientometrics, 75*, 37–50.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management, 35*(3), 401-420.

Chen, C. M., Cribbin, T., Macredie, R. and Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal of the American Society for Information Science and Technology, 53*(8), 678–689.

Chiusano, M. L., D'Agostino, N., Traini, A., Licciardello, C., Raimondo, E., Aversano, M., Frusciante, L. and Monti, L. (2008). ISOL@: an italian SOLAnaceae genomics resource. *BMC Bioinformatics, 9*(Suppl 2), S7.

Clark, W. C. (2001). Research systems for a transition toward sustainability. In Steffen, W., Jäger, J., Carson, D. J. and Bradshaw, C., editors, *Challenges of a changing earth: Proceedings of the global change open science conference, Amsterdam, 2001*, pages 197–200. Berlin.

Clark, W. C. (2007). Sustainability science: A room of its own. *Proceedings of the National Academy of Sciences, 104*(6), 1737–1738.

Clark, W. C. and Dickson, N. M. (2003). Sustainability science: The emerging research program. *Proceedings of the National Academy of Sciences of the United States of America, 100*(14), 8059–8061.

Clarke, T. (2002). Sustainable development: Wanted: scientists for sustainability. *Nature, 418*(6900), 812–814.

Crane, D. (1980). An exploratory study of kuhnian paradigms in theoretical high energy physics. *Social Studies of Science, 10*(1), 23–54.

Cronin, B. (2000). Semiotics and evaluative bibliometrics. *Journal of Documentation, 56*(4), 440–453.

Davidse, R. and Van Raan, A. F. J. (1997). Out of particles: Impact of CERN, DESY and SLAC research to fields other than physics. *Scientometrics, 40*(2), 171–193.

De Boer, Y., De Gier, A., Verschuur, M. and De Wit, B. (2006). Building bridges: Researchers on their experiences with interdisciplinary research in the netherlands. Technical Report Den Haag: RMNO.

De Bruin, R. E. and Moed, H. F. (1993). Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications. *Scientometrics, 26*(1), 65–80.

De Heer, J. and Card, S. K. (2004). DOITrees revisited: scalable, space-constrained visualization of hierarchical data. In *Proceedings of the working conference on Advanced visual interfaces* AVI '04, pages 421–424. New York, NY, USA.

De Mey, M. (1982). *The Cognitive Paradigm.* London: D. Reidel Publishing Company.

De Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149*(3683), 510–515.

DeLoache, J. S. and Marzolf, D. P. (1992). When a picture is not worth a thousand words: Young children's understanding of pictures and models. *Cognitive Development, 7*(3), 317–329.

Dillon, J. T. (1981). The emergence of the colon: An empirical correlate of scholarship.. *American Psychologist, 36*(8), 879–884.

Dillon, J. T. (1982b). In pursuit of the colon: A century of scholarly progress: 1880-1980. *The Journal of Higher Education, 53*(1), 93–99.

Dodds, S. (1997). Towards a 'science of sustainability': Improving the way ecological economics understands human well-being. *Ecological Economics, 23*(2), 95–111.

Eck, N. J. and Waltman, L. (2009). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Egghe, L. and Rao, I. K. R. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics, 25*(1), 5–46.

Egghe, L. and Ravichandra Rao, I. K. (1992). Classification of growth models based on growth rates and its applications. *Scientometrics, 25*, 5–46.

Egghe, L. and Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science, 51*(11), 1004–1017.

Egghe, L. and Rousseau, R. (2002). A proposal to define a core of a scientific subject: A definition using concentration and fuzzy sets. *Scientometrics, 54*(1), 51–62.

Elliot, J. A. (2006). *An introduction to Sustainable Development.* Abingdon: Routledge.

Figala, K. (2002). Newton's alchemy. In Cohen, I. B. and Smith, G. E., editors, *The Cambridge Companion to Newton*, pages 370–386. Cambridge: Cambridge University Press.

Fried, J., Heinemeier Hansson, D. and Linderman, M. (2009). *Getting Real: the Smarter, Faster, Easier Way to Build a Successful Web Application.* 37signals.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964–971.

Gambardella, A. and Torrisi, S. (1998). Does technological convergence imply convergence in markets? evidence from the electronics industry. *Research Policy, 27*(5), 445–463.

Garfield, E. and Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. New York: Wiley.

Garfield, E. and Sher, I. H. (1993). Keywords plus algorithmic derivative indexing. *Journal of the American Institute for Information Science, 44*(5), 298–299.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology, 5*(10), R80.

Gerdt, P., Kommers, P., Looi, C. K. and Sutinen, E. (2001). Cognitive technology 2001. In Beynon, M., Nehaniv, C. L. and Dautenhahn, K., editors, *LNAI 2117*, pages 233–247. Springer-Verlag, Berlin.

Gibbons, M., Limogenes, C., Nowotny, H., Schwartzman, S., Scott, P. and Trow, M. (1994). *The new production of knowledge.* Sage Publications.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821.

Goffman, W. (1971). A mathematical method for analyzing the growth of a scienfic discipline. *Journal of the Association for Computing Machinery, 18*(2), 1672–1673.

Goldbarth, A. (1997). ***!!!The battle of the century!!!***. *Kenyon Review, 19*(3-4), 143–147.

Gundersen, C. and Ziliak, J. P. (2004). Poverty and macroeconomic performance across space, race, and family structure. *Demography, 41*(1), 61–86.

Hammer, Linke, Wagner and Diefenbeck (1998). Organs from animals for man. *International Archives of Allergy & Immunology, 116*(1), 5–21.

Haniel, A. (2000). Organs at any price? on the question of alternatives to postmortem organ donation (examining ethical aspects of human organ cultivation and cloning). *Zeitschrift für Evangelische Ethik, 44*(4), 269–284.

Hartley, J. (2007). Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research, 29*(4), 553–568.

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J. and Wilson, S. (2008). What makes an article influential? predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.

Healey, P., Rothman, H. and Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233–251.

Hemmje, M. (1995). LyberWorld: a 3D graphical user interface for fulltext retrieval. In *Proceedings of the ACM/SIGCHI'95*, pages 417—418.

Hillis, D. M. and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology, 42*(2), 182 –192.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: the case of bioelectronics. *Scientometrics, 29*(3), 353–376.

Hinze, S. (1997). *Mapping of Structures in Science and Technology: Bibliometric Analyses for Policy Purposes.* PhD thesis, Leiden: Leiden University.

Hirsch, J. E. (2007). Does the h index have predictive power?. *Proceedings of the National Academy of Sciences, 104*(49), 19193–19198.

Hirst, G. (1978). Discipline impact factors: A method for determining core journal lists. *Journal of the American Society for Information Science, 29*(4), 171–172.

Hoede, C. (1994). On the ontology of conceptual graphs. Memorandum 1239, Enschede: Faculty of Applied Mathematics, University of Twente.

Hofstadter, D. (1985). *Metamagical Themas: Questing For The Essence Of Mind And Pattern.* New York: Basic Books.

Holmberg, J. and Sandbrook, R. (1992). Sustainable development: What is to be done?. In *Policies for a Small Planet*, pages 19–38. London: Earthscan.

Huff, A. S. (1990). *Mapping strategic thought.* Chichester: John Wiley & Sons.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264–323.

Janssens, F., Zhang, L., De Moor, B. and Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management, 45*(6), 683–702.

Järvelin, K., Ingwersen, P. and Niemi, T. (2000). A user-oriented interface for generalised informetric analysis based on applying advanced data modelling techniques. *Journal of Documentation, 56*(3), 250–278.

Kajikawa, Y. (2008). Research core and framework of sustainability science. *Sustainability Science, 3*(2), 215–239.

Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K. and Komiyama, H. (2007). Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science, 2*(2), 221–231.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters, 31*(1), 7–15.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Grübler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., Mooney, H., Moore, B., O'Riordan, T. and Svedin, U. (2001). Sustainability science. *Science, 292*(5517), 641–642.

Kates, R. W., Parris, T. M. and Leiserowitz, A. A. (2005). What is sustainable development? (cover story). *Environment, 47*(3), 8–21.

Kauffman, J. (2009). Advancing sustainability science: report on the international conference on sustainability science (ICSS) 2009. *Sustainability Science, 4*(2), 233–242.

Keitsch, M. M. (2010). Sustainability and science – challenges for theory and practice. *Sustainable Development, 18*(5), 241–244.

Kerr, A., Cunningham-Burley, S. and Amos, A. (1998). Drawing the line: an analysis of lay people's discussions about the new genetics. *Public Understanding Of Science, 7*(2), 113–133.

Kim, W. and Wilbur, W. J. (2001). Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology, 52*(3), 247–259.

Kirkby, J., O'Keefe, P. and Timberlake, L. (1995). *The Earthscan Reader in Sustainable Development*. London: Earthscan.

Klavans, R. and Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455–476.

Kohonen, T. (1998). Self-Organization of very large document collections: State of the art. In *Proceedings of ICANN98*, pages 65—74.

Komiyama, H. and Takeuchi, K. (2006). Sustainability science: building a new discipline. *Sustainability Science, 1*(1), 1–6.

Kostoff, R. N. (2002). Overcoming specialization. *BioScience, 52*(10), 937.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115–129.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Landesman, C. (1997). *An Introduction to Epistemology.* Oxford, UK: Blackwell Publishers Ltd..

Larivière, V., Archambault, É. and Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology, 59*(2), 288–296.

Larivière, V. and Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126–131.

Law, J. (1986). The heterogeneity of texts. In Callon, M., Law, J. and Rip, A., editors, *Mapping the dynamics of science and technology*, pages 67–83. London: The MacMillan Press Ltd.

Lee, K. N. (1993). *Compass and Gyroscope: Integrating Science And Politics For The Environment.* Washington: Island Press.

Lewison, G. (1999). The definition and calibration of biomedical subfields. *Scientometrics, 46*(3), 529–537.

Lewison, G. and Hartley, J. (2005). What's in a title? numbers of words and the presence of colons. *Scientometrics, 63*(2), 341–356.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40–54.

López-Illescas, C., Noyons, E. C. M., Visser, M. S., Moya-Anegón, F. D. and Moed, H. F. (2009). Expansion of scientific journal categories using reference analysis: How can it be done and does it make a difference?. *Scientometrics, 79*(3), 473–490.

Mai, J. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation, 57*(5), 591–622.

Matxain, J. M., Eriksson, L. A., Formoso, E., Piris, M. and Ugalde, J. M. (2007). Endohedral (X@ZniSi)i=4-160,+- nanoclusters, x = li, na, k, cl, br. *The Journal of Physical Chemistry C, 111*(9), 3560–3565.

Mayer, A. L. (2008). Strengths and weaknesses of common sustainability indices for multidimensional systems. *Environment International, 34*(2), 277–291.

Meadows, D. (1998). Indicators and information systems for sustainable development. *A report to the Balaton Group*, pp. 1–25.

Merton, R. K. (1968). The Matthew Effect in Science. *Science, 159*(3810), 56–63.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Moed, H., Bruin, R. D. and Van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Moed, H. F. (2005a). *Citation Analysis in Research Evaluation*, volume 9 of *Information Science and Knowledge Management.* Dordrecht: Springer.

Morillo, F., Bordons, M. and Gómez, I. (2003). Interdisciplinarity in science: a tentative typology of disciplines and research areas. *Journal of the American Society for Information Science and Technology, 54*(13), 1237–1249.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, cocitation and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, 56*(12), 1250–1273.

Mothe, J., Chrisment, C., Dkaki, T., Dousset, B. and Karouach, S. (2006). Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, Environment and Urban Systems, 30*, 460–484.

Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E. and Munoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics, 61*, 129–145.

Nobel Foundation (2001). Nobel laureates. . retrieved 2011-02-21.

Nordmann, A. (2004). Converging technologies — shaping the future of European societies. Technical Report EU HLEG Foresighting the New Technology Wave.

Novak, J. D. (1998). *Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations.* New Jersey: Lawrence Erlbaum Associates.

Noyons, E. C. M. (1999). *Bibliometric Mapping as a Science Policy and Research Management Tool.* PhD thesis, Leiden: Leiden University.

Noyons, E. C. M. (2004). Science maps within a science policy context. In Moed, H. F., Glänzel, W. and Schmoch, U., editors, *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&t Systems*, pages 237–255. Dordrecht: Kluwer Academic Publishers.

Noyons, E. C. M., Moed, H. F. and Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115–131.

Noyons, E. C. M. and Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics, 41*(1-2), 61–67.

Noyons, E. C. M. and van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68–81.

Nuutinen, J., Sutinen, E., Liinamaa, K. and Vanharanta, H. (2004). Woven stories as a tool for corporate strategy planning. In *Proceedings of the WBC2004*, pages 438–441. Lisbon.

O'May, K. (1966). Quantitative growth of the mathematical literature. *Science, 154*, 1672–1673.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441–453.

Ozmutlu, S. (2006). Automatic new topic identification using multiple linear regression. *Information Processing & Management, 42*(4), 934–950.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science, 50*(3), 242–253.

Parvulescu, A..*brather "brother" + indoeuropean etymology. *Indogemanische Forschungen, 101*, 99–106.

Pearce, D., Barbier, E. and Markandya, A. (1990). *Sustainable Development - Economics and Environment in the Third World.* Aldershot: Edward Elgar.

Persson, G. (1999). Mistra's role and some lessons learned in the first five years. *Policy Sciences, 32*(4), 323–326.

Peters, H. P. F., Braam, R. R. and Van Raan, A. F. J. (1995). Cognitive resemblance and citations relations in chemical engineering publications. *Journal of the American Society for Information Science, 46*(1), 9–21.

Phillis, Y. A. and Andriantiatsaholiniaina, L. A. (2001). Sustainability: an ill-defined concept and its assessment using fuzzy logic. *Ecological Economics, 37*(3), 435–456.

Plerou, V., Gopikrishnan, P., Gabaix, X. and Stanley, H. E. (2002). Quantifying stock-price response to demand fluctuations. *Physical Review E, 66*(2, Part 2).

Polanco, X., François, C. and Lamirel, J. C. (2001). Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. *Scientometrics, 51*(1), 267–292.

Porter, A. and Chubin, D. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3), 161–176.

Price, D. D. S. (1981a). The analysis of scientometric matrices for policy implications. *Scientometrics, 3*(1), 47–53.

Price, D. D. S. (1981b). The analysis of square matrices of scientometric transactions. *Scientometrics, 3*(1), 55–63.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* Morgan Kaufmann.

R (2005). Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

Rafols, I. and Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rinia, E., Van Leeuwen, T. N. and Van Raan, A. F. J. (2002). Impact measures of interdisciplinary research in physics. *Scientometrics, 53*(2), 241–248.

Rinia, E. J. (2007). *Measurement and evaluation of interdisciplinary research and knowledge transfer.* PhD thesis, Leiden: Leiden University.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2001a). Citation delay in interdisciplinary knowledge exchange. *Scientometrics, 51*(1), 293–309.

Rinia, E. J., Van Leeuwen, T. N., Bruins, E. E. W., Van Vuren, H. G. and Van Raan, A. F. J. (2002). Measuring knowledge transfer between fields of science. *Scientometrics, 54*(3), 347–362.

Roco, M. C. (2002). Coherence and divergence of megatrends in science and engineering. *Journal of Nanoparticle Research, 4*(1), 9–19.

Roco, M. C. and Bainbridge, W. S., editors (2003). *Converging Technologies for Improving Human Performance*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Ródenas-Torralba, E., Morales-Rubio, Á. and Guardia, M. D. L. (2006). Scientometric picture of the evolution of the literature of automation in spectroscopy and its current state. *Spectroscopy Letters, 39*, 513–432.

Rogatis, L. D., Montini, T., Casula, M. F. and Fornasiero, P. (2008). Design of Rh@Ce0.2Zr0.8O2–Al2O3 nanocomposite for ethanol steam reforming. *Journal of Alloys and Compounds, 451*, 516–520.

Rosenberg, N. (1976). *Perspectives on technology*. Cambridge: Cambridge University Press.

Ruiz-Baños, R., Bailón-Moreno, R., Jiménez-Contreras, E. and Courtial, J. (1999). Structure and dynamics of scientific networks. part i: Fundamentals of the quantitative model of translation. *Scientometrics, 44*(2), 217–234.

Saeed, J. I. (2003). *Semantics*. Malden (MA): Blackwell.

Sagi, I. and Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science, 34*(5), 687, 680.

Sangermano, M., Perruchas, S., Gacoin, T. and Rizza, G. (2008). Synthesis of Au@SiO2 Core/Shell nanoparticles and their dispersion into an acrylic photocurable formulation: Film preparation and characterization. *Macromolecular Chemistry and Physics, 209*(22), 2343–2348.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics, 59*(3), 425–465.

Schwechheimer, H. and Winterhager, M. (1999). Highly dynamic specialities in climate research. *Scientometrics, 44*(3), 547–560.

Sharachchandra, M. L. (1991). Sustainable development: A critical review. *World Development, 19*(6), 607–621.

Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation, 28*(11), 758–775.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to Non-Parametric regression curve fitting. *Journal Of The Royal Statistical Society Series B-Methodological, 47*(1), 1–52.

Simar, L. (1992). Estimating efficiencies from frontier models with panel data: A comparison of parametric, non-parametric and semi-parametric methods with bootstrapping. *Journal of Productivity Analysis, 3*, 171–203.

Sladkov, V., Guillou, V., Peulon, S. and L'Her, M. (2004). Voltammetry of tetraalkylammonium picrates at waternitrobenzene and waterdichloroethane microinterfaces; influence of partition phenomena. *Journal of Electroanalytical Chemistry, 573*(1), 129–138.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799–813.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*, 595–610.

Small, H. E. and Upham, S. P. (2009). Citation structure of an emerging research area on the verge of application. *Scientometrics, 79*(2), 365–375.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327–340.

Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes, 26*(1), 90–102.

Solomon, R., Sandborn, P. A. and Pecht, M. G. (2000). Electronic part life cycle concepts and obsolescence forecasting. *IEEE Transactions on Components and Packaging Technologies, 23*(4), 707–717.

Sowa, J. F. (1983). *Conceptual structures: Information processing in mind and machinE*. Reading, Mass.: Addison-Wesley.

Srinivas, V. and Shekar, B. (1997). Applications of uncertainty-based mental models in organizational learning: A case study in the indian automobile industry. *Accounting, Management and Information Technologies, 7*(2), 87–112.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy, 25*(5), 701–721.

Stanley, H. E., Amaral, L. A. N., Canning, D., Gopikrishnan, P., Lee, Y. and Liu, Y. (1999). Econophysics: Can physicists contribute to the science of economics?. *Physica A: Statistical Mechanics and its Applications, 269*(1), 156–169.

Stockwell, P., Colomb, R. M., Smith, A. E. and Wiles, J. (2009). Use of an automatic content analysis tool: A technique for seeing both local and global scope. *International Journal of Human-Computer Studies, 67*(5), 424–436.

Tabah, A. N. (1999). Literature dynamics: studies on growth, diffusion and epidemics. In Williams, M. E., editor, *Annual Review of Information Science and Technology*, pages 249–286. Information Today Inc.

Takeda, Y. and Kajikawa, Y. (2009). Optics: a bibliometric approach to detect emerging research domains and intellectual bases. *Scientometrics, 78*(3), 543–558.

TC 159/SC 4 (2000). Human-centred lifecycle process descriptions. *ISO standard 18529.*

Thorndyke, P. W. and Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology, 14*(4), 560–589.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.

Tijssen, R. J. W. (1992). *Cartography of Science: Scientometric Mapping with Multidimensional Scaling Techniques.* PhD thesis, Leiden: Leiden University.

Tijssen, R. J. W., Hollanders, H., Van Leeuwen, T. N. and Nederhof, A. J. (2008). Science and Technology Indicators 2008. Netherlands Observatory of Science and Technology. Technical Report Leiden: NOWT.

Tseng, Y., Lin, Y., Lee, Y., Hung, W. and Lee, C. (2009). A comparison of methods for detecting hot topics. *Scientometrics, 81*(1), 73–90.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, Connecticut: Graphics Press.

Uhlenbeck, G. E. and Goudsmit, S. (1925). Ersetzung der hypothese vom unmechanischen zwang durch eine forderung bezüglich des inneren verhaltens jedes einzelnen elektrons. *Die Naturwissenschaften, 13*, 953–954.

Upham, S. P. and Small, H. E. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics, 83*(1), 15–38.

Urata, H. (1990). Information flows among academic disciplines in Japan. *Scientometrics, 18*, 309–319.

Van den Besselaar, P. and Heimeriks, G. (2001). Disciplinary, multidisciplinary, interdisciplinary: Concepts and indicators. In *Paper the 8th ISSI Conference for Scientometrics and Informetrics, Sydney* .

Van Eck, N. J. and Waltman, L. (2007). VOS: a new method for visualizing similarities between objects. In Decker, R. and Lenz, H. J., editors, *Advances in Data Analysis*, pages 299–306. Berlin, Heidelberg: Springer Berlin Heidelberg.

Van Eck, N. J., Waltman, L., Noyons, E. C. M. and Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics, 82*(3), 581–596.

Van Leeuwen, T. N. (2004). *Second generation bibliometric indicators – the improvement of existing and development of new bibliometric indicators for research and journal performance assessment procedures.* PhD thesis, Leiden: Leiden University.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.

Van Raan, A. F. J. (1999). Advanced bibliometric methods for the evaluation of universities. *Scientometrics, 45*(3), 417–423.

Van Raan, A. F. J. (2000). On growth, aging and fractal differentiation of science. *Scientometrics, 47*(2), 347–362.

Van Raan, A. F. J. (2008). Bibliometric statistical properties of the 100 largest european research universities: Prevalent scaling rules in the science system. *Journal of the American Society for Information Science and Technology, 59*(3), 461–475.

Van Raan, A. F. J., Van Leeuwen, T. N., Visser, M. S., Van Eck, N. J. and Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. *Journal of Informetrics, 4*(3), 431–435.

Vibert, N., Ros, C., Bigot, L. L., Ramond, M., Gatefin, J. and Rouet, J. (2009). Effects of domain knowledge on reference search with the PubMed database: An experimental study. *Journal of the American Society for Information Science and Technology, 60*(7), 1423–1447.

Voutilainen, A. (1995). NPtool, a detector of english noun phrases. *cmp-lg/9502010.*

Waltman, L., van Eck, N. J. and Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*(4), 629–635.

Waltman, L., Van Eck, N. J., Van Leeuwen, T. N., Visser, M. S. and Van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47.

Weinberg, A. M. (1972). Science and trans-science. *Minerva, 10*(2), 209–222.

White, H. D. and McCain, K. W. (1997). Visualization of literatures. In Williams, M. E., editor, *Annual review of information science and technology (ARIST)*, pages 99–168. Information Today, Inc. for the American Society for Infomation Science.

White, H. D. and McCain, K. W. (1998). Visualizing a discipline: An author cocitation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

Williams, J. G., Sochats, K. M. and Morse, E. (1995). Visualization. In Williams, M. E., editor, *Annual review of information science and technology (ARIST)*, pages 161–207. Information Today, Inc. for the American Society for Infomation Science.

Wouters, P. (1998). The signs of science. *Scientometrics, 41*, 225–241.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11.ACM

Young, M. D. (1996). Cognitive mapping meets semantic networks. *Journal of Conflict Resolution, 40*(3), 395.

Zitt, M. and Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information processing & management, 42*(6), 1513–1531.