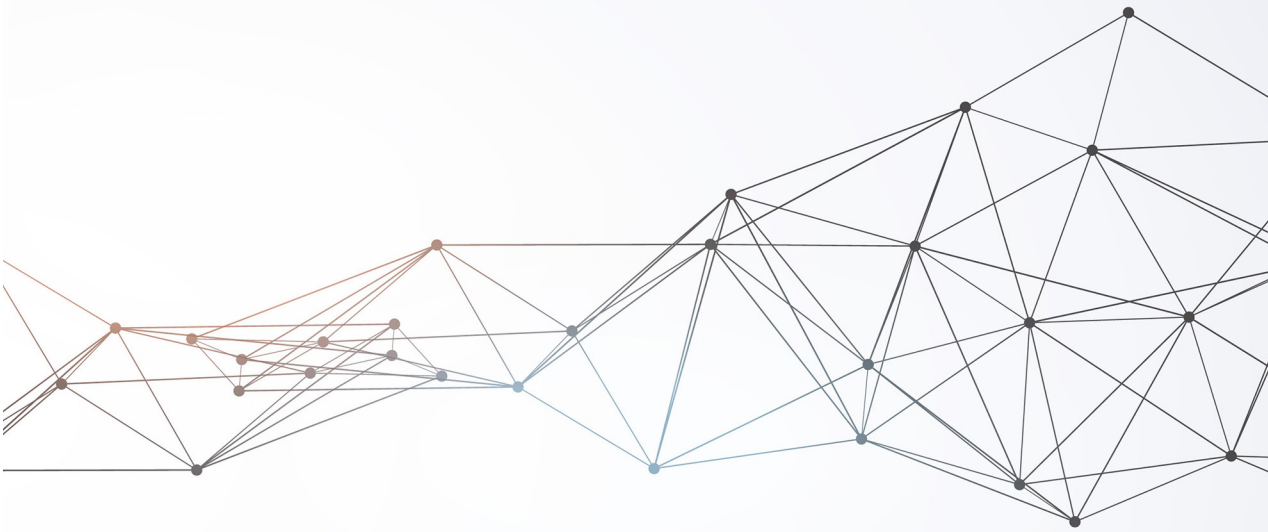
TANULMÁNYOK A TUDOMÁNYELEMZÉS MAI GYAKORLATÁBÓL

STUDIES FROM THE PRESENT-DAY PRACTICE OF SCIENTOMETRICS



Sándor Soós

ENHANCED KNOWLEDGE DISCOVERY IN SCIENTIFIC COMMUNITIES

ENHANCED KNOWLEDGE DISCOVERY IN SCIENTIFIC COMMUNITIES

TANULMÁNYOK A TUDOMÁNYELEMZÉS MAI GYAKORLATÁBÓL

STUDIES FROM THE PRESENT-DAY PRACTICE OF SCIENTOMETRICS

2

SOROZATSZERKESZTŐ:

SERIES EDITOR:

SOÓS SÁNDOR

Enhanced knowledge discovery in scientific communities

via the integration of science mapping methods and its "proof of concept" application to complex research problems

Sándor Soós, PhD

2023

Budapest, 2024

The book is based on a work supported by the
János Bolyai Research Scholarship of the Hungarian Academy of Sciences.
Additionally, it serves as a habilitation thesis that has been reviewed and accepted as part of the habilitation process at the Faculty of Education and Psychology, Eötvös Loránd University (ELTE).

BOLYAI JÁNOS
KUTATÁSI
ÖSZTÖNDÍJ

The volume is published on the occasion of the 200th anniversary of the foundation of the Library and Information Center of the Hungarian Academy of Sciences, and the festive program series "MTA200".
Published with the support of the Hungarian Academy of Sciences.

MTA 200 ÉVES
A MAGYAR
TUDOMÁNYOS
AKADÉMIA

Responsible Publisher: the Director General of the Library and Information Center of the Hungarian Academy of Sciences
Series Editor: Sándor Soós
Typography and typesetting: Viktória Vas
Printed and bound by Prime Rate Kft.
Cover image source: Freepik.com

ISBN 978-615-6792-03-7
DOI 10.36820/tudomanyelemzes.2023.2
ISSN 2677-1683

Table of Contents

Introduction: representing and modelling scientific knowledge

The present book is the result of an extensive research aiming the advance of a specific methodological framework developed for modelling and formally representing scientific knowledge – based on large-scale, empirical data on scholarly discourses. This framework is generally referred to as *science mapping* (SM), or *representing scientific knowledge* (Chen & Song, 2017) The main objective of the reported research was twofold: First, to further elaborate on selected instruments that, by their combination and integration, have an increased capability to model the various aspects of the cognitive structure of research problems and knowledge domains, including their causal interactions that shape the body of knowledge associated in the respective domains of science. Second, to provide a "proof-of-concept" for our work, a specific knowledge domain was addressed via the proposed science mapping instruments in a detailed case study, both to demonstrate their analytic capacity and also to analyze their validity. This scientific domain is the so-called *species problem*, a widely interdisciplinary discourse at the intersection of biological systematics, evolutionary and theoretical biology, the philosophy of science, cognitive psychology and cognitive anthropology. With such complexity, this research problem is an ideal candidate to test the added value of our proposed knowledge representation models and measures. In this preparatory chapter we first contextualize science mapping (SM) as a methodological framework, then briefly introduce the species problem, i.e. the case study we use for the purposes of validation and demonstration, finally we turn to the overview of the structure of this book.

The conceptualization of scientific knowledge in science mapping

Science Mapping as a formal methodology of knowledge representation is best conceived as being instrumental for social epistemology as a theoretical framework through mathematical modelling (Goldman, 2011). Social

epistemology aims to account for scientific knowledge as the product of the scientific community, and seeks to explain its characteristics with reference to community-level cognitive interactions instead of cognitive processes (exclusively) bound to the individual level. Originating from the philosophy of science, social epistemology quickly enrolled models and empirical support from cognitive science through such concepts as "distributed cognition" (Thagard, 2012), modelling (scientific) knowledge as the emergent outcome of cognitive processes distributed among interacting agents, being also (in part) external to individuals – as contrasted with the classical or computational school of cognitive science where knowledge was conceived via formal models representing individual cognitive processes. This distributed, "supraindividual" nature of scientific knowledge, not fully attributable to individuals also gives room to (or even implies) the "body" of knowledge resulting from the interaction of agents having an inherent variability. In terms of the evolutionary models of science (Hull, 1990), this body of knowledge is constituted of its individual variants distributed in the scientific community, i.e. a more-or-less diverse population of theories, concepts, representations, models that co-exist as a(n evolving) paradigm, research programme or tradition. Hence the term coined by one of the most prominent figures in science mapping (Chen, 2013), who dubbed the underlying construct subjected to representation "latent domain knowledge". This term nicely captures that the mission of science mapping is to model a construct of knowledge that, despite being explicit or formal (as contrasted with implicit or tacit), can only be reconstructed through the empirical exploration of its variants and their interrelations across the scientific community. In sum, for representing scientific knowledge (SK) framed within social epistemology, we have to use models accounting for all these intertwined characteristics, i.e. SK being (1) community-level, (2) distributed, (3) external to individuals, (4) existing in variants (5) explicit and formal, but still latent. As we shall see in the upcoming sections, science mapping is a viable analytical framework to operationalize such a construct of scientific knowledge.

The operationalization of scientific knowledge in science mapping

The abovedescribed conceptualization is best summarized in Bird (2010) making a case for scientific knowledge that emerges from the cognitive interactions within the scientific community (or, rather, communities), but being distinct from the aggregation of individuals' knowledge. Most importantly for our present purposes, as Bird argues, it is encoded in scientific communication, hence embodied and represented through its formal channels, scholarly publications, journals, books etc. Science mapping instruments therefore operationalize the construct of knowledge through the study of observational indicators of scientific communication, and rely on empirical data drawn from large-scale citation and publication databases. Through these indicators, the models and related measures aim to operationalize the relevant cognitive interactions and interelations between the actors of the community to arrive at a formal representation of the latent community-level knowledge. Such indicators involve citation relations between publications, authors, journals etc. conveying knowledge flow, knowledge sharing and distribution; co-authorships as socio-cognitive interactions conveying the formation of cognitive communities; textual data and metadata for modelling the conceptual organization of the domain through the statistical analysis of their relations, also accounting for the variability within the discourse – just to name the most frequent examples.

The methodological character of science mapping (SM)

Science mapping is best characterized within the toolbox of formal social epistemology. The latter shares an important methodological feature with classical cognitive science as it employs formal – computational models in explaining and exploring cognitive behavior and knowledge. For science mapping, these models for the representation of scientific knowledge are usually network models, hence the analysis of these knowledge structures

is typically provided in terms of network science, the analysis of complex networks adopted and elaborated for knowledge networks. In fact, it was one of the most prominent figures of contemporary network science, Barabási Albert-László who, with his colleagues, repositioned science mapping as a branch of network science reviving its original name, "the Science of Science" (Wang & Barabási 2021). However, it should be noted that SM as a methodology has further historical roots and methodological aspects. Since the methods rely on large-scale bibliographic and citation data and metadata of scholarly documents, they are often considered as part of bibliometrics, scientometrics and informetrics, or quantitative science studies, in general. This attribution is also sound historically, since early approaches that determined the development of the field (dating back to the 1970s–1980s) came from information science. From the viewpoint of knowledge representation, however, these informetric and network-analytic methods serve as a toolkit of explorative instruments. Perhaps the most adequate approach has been provided by Hjørland, who demonstrated (Hjørland, 2013) that citation analysis, as a family of models in science mapping, is a field of KO, i.e., Knowledge Organization, classifying bibliometric maps as knowledge organization systems (KOs). Most importantly for our work, the argument shows that such maps are capable of reconstructing *causal links* in the community-level development of a body of knowledge, so that the conceptual organization of, e.g, a research problem can be empirically grounded instead of subjective or individual interpretations.

The parallel objective: a complex case study of an interdisciplinary research problem

The primary focus of the research presented in this book is the advancement of the methodology of knowledge representation and science mapping, in particular. Nevertheless, its application to a carefully chosen scientific discourse gains almost equal weight in the discussion of our results and, in

fact, throughout the entire book. This is beacause we intended to provide a fully elaborated case study as a "proof of concept" for our methods which serves multiple purposes. First of all, (1) it demonstrates the operation and the practical value of the methods. Second, (2) it provides an opportunity to validate these instruments not only on the quantitative but also on the qualitative level, against the expert reviews of the discourse. Third (3) the case study as a research outcome can also be considered a contribution on its own, since, to the author"s knowledge, this is the first attempt to empirically and systematically discover the organization of this discourse.

The case study in question is that of the scientific discourse called "the species problem". In the history of biology, one can found some big questions and controversies that both drive the development of the discipline and seem to be unresolved (or, even, unresolvable) to date, at the same time. Such an ever-green controversy, originating from the ancient prehistory of the life sciences, is the *Species Problem*. The species problem, simply put, is the debate on the valid, scientifically sound concept – definition – of biological species. Simple as it seems, the standard story of the species problem begins with the work of Aristotle and Plato, leads to the work of Darwin, and penetrates biological systematics and evolutionary biology throughout the XX. century. In addition to the vast historical record of the species debate in the history of science, "secondary" literature focusing on the causes, drives, structure and, above all, the durability – that is, the puzzling nature – of the discourse has also been flourishing in the last decades of the XX. century. The loci of this discourse is primarily the philosophy of science (of biology, as a distinct field), the "internalist" historiography of science, and various works from theoretical biology. However, various fields of supraindividual biology (systematics, evolutionary biology, genetics and molecular biology, ecology, microbiology etc.) have also heavily contributed to the debate. What's more, the persistence of the problem has lately been connected to the psychological constraints of human cognition in cognitive science

(c.f. Lopez et al., 1997), so that cognitive psychology and cognitive anthropology have also been drawn to this knowledge domain. The interaction of versatile fields resulted in a deeply interdisciplinary domain, ideal for testing our knowledge representation methodology for its capabilities to discover and formally reproduce the complex knowledge structures along with their causal interactions. This latter criterion, the discovery of causal interactions is key to our approach: it enables these instruments to validate the narrative or "expert" reviews as well (and not just vice versa), that is, to confirm or disconfirm the explanations on how interdisciplinary relations – in particular, that of philosophy and biology – shaped our knowledge on species (as discussed in the chapter with the narrative review of the species problem), resulting in a certain cross-validation of the two.

Research questions and the organization of the book

As a summary of this brief introduction, we can formulate the main objective of the presented research as the elaboration of a knowledge representation toolkit by the combination and further advancement of science mapping methods aimed at the cognitive organization of scientific domains. In addition, to validate this framework, a detailed case study of a highly complex and interdisciplinary discourse will be analyzed via the proposed methodology, the so-called "Species Problem". With regard to the representational capacity and the qualitative (and, where appropriate, quantitative) validation of our instruments, we focus on the following "diagnostic" research questions:

RQ1. *Case study results*. What does the application of the proposed methodology reveal about the "real" (and latent) empirical structure of the discourse on the species problem (in the selected time period)? In particular what conceptual systems, research traditions, research fields engage in causal itneractions, when and how, in the formation of the discourse?

RQ2.a *Qualitative validation.* Can the proposed, integrated methodology be validated by reproducing the key conceptual interactions and dynamics shaping the discourse according to the expert-based literature review? Is the "standard story" on the interdisciplinary problem of species confirmed by the instruments and vice versa?

RQ2.b *Quantitative validation*. Do the proposed methods outperform existing or conventional science mapping methods in quantitative comparisons? – this subquestion will be discussed along with the Methods section, along with the detailed introduction of the proposed methods in empirical and theoretical comparisons, whereever appropriate.

RQ3. *Instrumental value*. Can the proposed methodology identify the causal interactions between fields and disciplines, yielding actual knowledge integration? Can it support the hypothesis that the discourse is truly interdisciplinary, where one field has a measurable impact on the other? (in our case, that of the philosophy of science on biological conceptualizations)?

To proceed with the introduction and quantitative study of the instruments as well as with the case study and qualitative results, the rest of the book is organized as follows:

- *Chapter 1* will introduce the case study with a narrative literature review of the species problem, which is based on standard reconstructions of the discourse. This will serve as a reference point against which the new instruments will be qualitatively validated through the empirical results of their application (whether they can reproduce and confirm the key causal factors in shaping the cognitive structure).
- *Chapter 2* will cover the data collection process for representing the scholarly discourse in our case study, the Species Problem.
- *Chapter 3* will delineate the discourse with empirical methods and set up its base model as a citation network upon which our proposed models and instruments will operate.

- *Chapter 4* will briefly overview the elementary methods of science mapping providing a concise taxonomy of the existing tools and forming the basis for our methodological work.
- *Chapter 5* is the place to present the methodological results from developing three, interrelated science mapping methods (*age-sensitive bibliographic coupling*, *multidimensional science maps* and *dynamic overlay maps*) addressing different aspects of the causal interactions (based on citation relations) in the conceptual development of the discourse. Where appropriate, Q2b is addressed by performance comparisons between the proposed and pre-existing instruments using the case study.
- *Chapter 6* will present a distinct methodological proposal addressing the latent conceptual organization of the discourse (*topic overlay maps*) and its performance againts conventional method (Q2b)
- *Chapter 7* is a multi-purpose section of the book. First, it will lay out the results from the application of the proposed methods on the species discourse, that is, the aspects and levels of its conceptual – causal structure revealed by the individual instruments (Research Question RQ1). The integration of these methods also belongs in this chapter by demonstrating, through the results, that applying them in tandem reveals those relevant levels and aspects where the latent formation of the discourse takes place. Also, the parallel presentation and discussion of the formal models of the debate will involve a constant comparison with the narrative review, the "standard story", for the purpuses of qualitative validation (hence RQ2).
- *Chapter 8* The final chapter, beyond conveying concluding remarks, will mainly revisit the remaining research question (RQ3), concerning whether our models (and results) do support the hypothesis on actual knowledge integration within this domain.

Chapter 1: The standard version of the Species Problem: a narrative review

In this chapter we (re)construct a narrative review of the Species Problem with the primary aim to provide a reference for both the performance assessment and the validation of the proposed instruments. In science mapping studies it is a common and consensual procedure to use the evaluation of field experts for the validation of the results (Gläser, 2020), that is, to contrast the the empirical and formal model of a scientific domain with the insights of the researchers, experts or "practicioners" within that domain. In our case, the "expert viewpoint" is conveyed by the narrative review below, an approach that has two justifications. First, the review builds on paradigmatic overviews of the knowledge domain and refers to the standard sources (key authors and seminal papers) identified by these overviews. Hence the label "standard version" that we apply for this review. Second, the selection of these sources is underpinned by a previous work of the author of this book that included an extensive research on the history of the Species Problem (Soós, 2008).

The Species Problem is usually characterized as being composed of two, interrelated research questions: (1) what is the valid definition for the concept or category of species, and (2) what is the proper methodology to delineate individual species called species taxa (Hey, 2011b).

The standard historiographic literature on the two issues are problematic, beyond the known issues of reconstructing histories. Most reconstruction within the history and philosophy of science provide a view on the historical development of the SP in terms of the modern philosophy of biology. One reason for that is that these reconstructions have long been drawn by philosophers and theoreticians of biology, and are conveying the views and theoretical sense of those scholars and analysts (McOuat, 1996). On the other hand, the discourse itself is deeply embedded in the development of modern systematics and evolutionary biology, and therefore goes hand-in-

hand with the history of whole and complex fields of modern biology. This makes the identification of relevant issues highly difficult, let alone the relevant historical or causal network between those issues.

Given all that complexities, the paradigmatic (and preliminary) story of the SP can be well told, in a milestone-based fashion. This story pinpoints two watersheds, and, consequently, three main periods for the SP. The first period (starting from greek philosophers, Plato and Aristotle) is marked by the work of Darwin, ending in 1859 as the famous year of publishing *The Origin of Species*. The next turn, taking place already in the XX. century, is dated to the 1940's, and referred to as the *evolutionary synthesis* (hereafter: Synthesis). The term points toward a period of revolutionary science forming the synthesis of classical genetics and the Darwinian theory of evolution, which resulted in a paradigm change in biology. The new paradigm, the *evolutionary paradigm*, that underlies modern biological science, characterizes the third period of the story. The Synthesis is attributed to two main figures within the history of biology: the pioneering work of Theodosius Dobzhansky and Ernst Mayr placed the controversy on species and the species concept in a radically new context, creating the basis for the modern view of species taxa. The third period (from the 1940's up to the present) can be conceived as the proliferation of this modern controversy on species.

In what follows, a brief overview is given to the "classical" versus the "modern" species problem, as we intend to use these terms. By "classical" we refer to the developments preceeding the Synthesis, while "modern" signifies post-Synthesis developments. The overview is in accord with the views of most authors on the history of the SP.

The classical view and species problem

The classical species problem can, historically, be identified in light of the huge amount of observations, and their theoretical explanation communicated by Darwin in the *Origin of Species.* Pre-Darwinian natural history, up from Aristotle, treated biological species as so-called *natural kinds* or *types*, embodied by specimens (individuals belonging to a species). According to Mayr (1982), this view is rooted back to Plato's philosophy, and calls this view the *typological species concept* (or, rather, species notion). The typological concept is, indeed, also the basis for the classical taxonomy of Carl von Linné introduced in the 18. century, as often reconstructed from the taxonomic principles layed down in his Systema Naturae (Linné, 1735). According to these principles, a species is a class of organisms exclusively sharing certain characteristics or properties (so that those properties are shared by all and only the specimens belonging to a species). These properties are called *essential*, by which any species can be delimited, unambiguously. Essential properties, therefore, serve as the definition for the respective species (taxon): they provide a necessary and sufficient set of criteria for assigning any organisms to their species. This *essentialist* view inherent to both Linné and the whole pre-Darwinian philosohy of nature is rooted in metaphysics. The "natural sysem" in Linné's eyes, the uncovering of which is the ultimate goal of any taxonomy or systematics, is the divine order of the universe: including all taxonomic units, like species. The creationistic idea of natural types further implies that species – that is, essences – are fixed and unchangeable. The typological species concept is, consequently, further characterized with *fixism* beyond essentialism.

Darwin's *Origin of Species* entered directly into the context described above. Both pillars of his classic work, *víz.* the reconstruction of "descent with modification" as an evolutionary process, and the elaboration of a general background theory explaining this process were based on a detailed, extremely well-documented observational study of species – in

the spirit of nineteenth century naturalism. The key Darwinian observation, in this respect, was that species – that is, groups categorized as species by contemporary naturalists – are, in reality, made up by individual variations. There is a clear diversity not only between species taxa, but within a species taxon as well. Darwin demonstrated, in the possession of a huge amount of evidence from the field, that the distribution of individual characteristics within any species is different from that of "essential properties", so that one cannot find a certain set of traits for a species taxon capable of identifying all and only the members of that taxon. A corollary of this finding is that species boundaries are actually vague and fuzzy, or, at least, cannot be objectively defined. On the other hand, intra-specific categories (such as subspecies, variates etc.) are also implicated, as telling apart these intra-specific taxa from each other also necessarily fails by the same principle.

According to the standard story, these findings led to the Darwinian view of species, stating that the concept is merely a means of convenience for the naturalist, allowing for scholarly communication, the description of the subject under study. The *Origin of Species*, therefore, often claimed paradoxical (as to its title): one consequence of the proposed theoretical framework for the explanation of where species come from is the destruction (or, rather, deconstruction) of the concept of species (Beatty, 1985). It is important to note that the so-called *nominalist view*, according to which species exist only as purely mental/linguistic categories, is much older than the *Origin.* It is the view attributable already to Buffon, whose eighteenth-century species definition was based on relations (kinship) among – rather than traits of – organisms, such that it can be considered as an intellectual precursor of the modern *Biological Species Concept*. The main proponent of the latter, Ernst Mayr calls this concept the *nominalistic species concept* (Mayr, 1982). However, some historians of biology doubts that Darwin would have been a true proponent of nominalism: Beatty, for example, argues that the Darwinian concept is much more sophisticated, and (anticipating the modern debate)

distinguishes between species as a *category* and as *taxa* (individual species): for Darwin, individual species are very realistic, formed and shaped by the natural processes he conjectured and described; what, on the other hand, is not real, is the *category of species*, which gains its reality from a definition, a set of criteria to delineate taxa (Beatty, 1985, Stamos, 1996).

The modern Species Problem

The conceptual framework established by Darwin, that addressed multiple interrelated issues including the nature of species, the theory of their descent and relations, and, most importantly, the causes accounting for their modification and – in that sense, evolution – has been revived in the 1940's, mostly due to the pioneering work of Theodosius Dobzhansky. In the light of early developments in classical genetics, Dobzhansky filled the Darwinian framework with new content from twentieth century experimental biology. The breakthrough is usually attributed to the Synthesis, incorporating Mendelian genetics into the Darwinian theory of evolution. In particular, Dobzhansky provided evidence from experimental genetics supporting the theory of natural selection, as the key process proposed by Darwing behind evolutionary change. The revival of evolutionary theory directly re-contextualized the related concepts, first and foremost the concept of species. Fed by this new context, the so-called Biological Species Concept (BSC) was introduced by Ernst Mayr, the now-historical figure of evolutionary systematics. According to this definition, species taxa are groups of organisms that are each reproductively isolated from other such groups. In other words, a species is made up of individuals – or, rather, populations – actually or potentially capable of (successful) interbreeding (Mayr 1963a). The fundamental novelty in this definition, compared to previous notions, that it goes beyond a simple set of criteria to delineate species, and offers also an explanation for the formation and existence of species taxa. The BSC, by its definition, targeted those biological factors that account for the distribution

of organisms into real, separate, units of nature (Mayr refers to this feature as the main motiv for the term “biological”). The explanatory power of this concept can be devised from the synthetic theory of evolution: in the context of this theory, the BSC category is equvivalent to the that of the “maximal Mendelian population” coming from Dobzhansky: within such populations gene flow is unrestricted, while impossible between them. Mayr used this factor to explain, by means of the BSC, the coherence and distinctiveness of species taxa: via reproductive isolation species act as *protected gene pools* each, which implies a relative and, at an evolutionary timescale, transitional, uniformity, that impress the observer as if species were unchangeable, fixed essences. To put it another way, the BSC was the first theoretical approach, that aimed at integrating the concept into the fabric of evolutionary theory in a scientific way.

The new concept of species, as is often emphasized by its proponent, differed from the classic view in various other ways as well (Mayr, 1996). For one thing, since the BSC did not rely on the assumption of essential properties to individuate species taxa, the concept was perfectly in accord with the Darwinian idea of species composed of individual variations. This view, also attributable to Mayr, is referred to as *population thinking*, as contrasted with *typological thinking*, the latter being predominant in previous eras.

Due to both the paradigm shift induced by this concept, and the far-reaching consequences of adopting the BSC, a vivid debate has emerged upon this new complexities of defining species. The modern species problem (at least its most dominant aspect within the history of biological science) can be framed and sketched out as the critical reception of the BSC throughout the systematics community. Most importantly, upon exploring more and more features and implications of this seemingly simple construct, theoreticians came up with more and more alternative concepts. These responses can be sorted into three or four families of species concepts, upon the distinctive types of criteria employed to build the category of species (Ereshefsky, in

press, Brigandt, 2003). Based on such a typology, we can distinguish between the Phenetic Species Concept (as the sole member of its family), ecological species concepts ("Ecospecies"), and phylogenetic species concepts ("Phylospecies").

Chronologically, the first construct emerging from the follow-up debate of the BSC was the Phenetic Species Concept. Its proponents, Sokal and Crovello challenged the BSC in two main respects (Sokal–Crovello, 1970): On one hand, they objected that the BSC does not provide operational means for the delineation of species taxa. The practice of applying the BSC in the field, as they pointed out, is still relying on observable similarities, that is, phenetic traits of candidate specimens. More importantly, they argued that the BSC was also theoretically unjustified, since the criterion of the capacity to interbreed is a theory-driven, *a priori* constraint, which tends to hinder the natural patterns of biodiversity. In order to obtain an empirical mapping of how the living world is organized, one that is free of restrictive theoretical commitments, Sokal and his colleagues voted for the methods of *numerical taxonomy*. As a school of systematics, numerical taxonomy relied on purely observational information, and offered computational models from statistics to sort organisms into taxa from large-scale records of phenetic data. In sum, the Phenetic Species Concept (PSC) adhereing to this tradition, challenged the theory-driven BSC and also represented the opposite methodological standpoint, excluding any taxonomic principle utilizing biological forces, processes or factors considered as predominant in maintaining species taxa.

The special focus placed by the BSC on interbreeding and isolation, as the main and definitive forces underlying the reality of species, has been questioned by several other directions of biology as well. Narrowing the concept of species down to interbreeding communities was found to be too restrictive for dealing with many practical cases in systematics, such as for asexual taxa where sexual reproduction is actually missing (like in microorganisms), or in cases where clearly conspecific populations, for natural reasons, remain

isolated. In the light of numerous real-life counterexamples, a different set of biological factors were proposed by Ehrlich, Raven and Van Valen, acting as definitive causes for the existence of species (Ehrlich–Raven, 1969, Van Valen, 1976). These authors argued that uniformity among conspecific organisms is attributable to being exposed to a *common selection regime*, so that coherent groups, i.e. species can emerge as a result. Van Valen described this view in terms of ecology with his *Ecological Species Concept* (ESC), claiming that species are groups of organisms – more precisely, lineages of organisms – that "occupy the same adaptive zone or ecological niche". It follows that the so-called ecological species are individuated by the joint distribution of their ecological parameters, that may overlap, but should at least minimally differ for different species. Bing adapted to such zones of environmental factors also explains the relative stability of species as well.

Frustrated by the vivid but rather unfruitful debate on the true biological factors accounting for the duration and separability of taxa, many theoreticians have chosen a different direction to grasp the species category. Leaving behind causes and processes as an inventory for constructing a definition, these authors rather turned – again – to recognizable patterns of nature to approach the ideal concept. As contrasted to previous (or, contemporary) pattern-oriented approaches, such as numerical taxonomy and the PSC, pattern-based proposals aimed at integrating the theories, biological mechanisms and causes highlighted by other species concepts (such as the BSC or ESC), as being in line with, or building on the merits of those concepts, but without distinguishing, or even referring any of these factors as definitive of species. This was the context of the introduction of three distinct construct, the so-called *evolutionary*, the *phylogenetic*, and the *cladistic* species concept. The often cited *Evolutionary Species Concept* credited to Simpson and Wiley (EvSC, Wiley, 1978) defines species as a "lineage of ancestral descendant populations which maintains its identity from other such lineages and which has its own evolutionary tendencies

and historical fate". A follow-up concept, also centered around lineages of populations, was the *Phylogenetic Species Concept* (PhSC, Cracraft, 1983), equating species with those lineages that share a unique combination of so-called novel characteristics (realtive to other lineages). In a different wording, a species taxon is claimed to be the smallest "diagnosable cluster" of organisms that are linked by the ancestor–descendant relation. The pattern-based family of definitions also incorporates the *Cladistic Species Concept* (CSC, Mishler–Donoghue, 1982), obtaining its name after cladistics, originated in a german school of twentieth-century systematics. Cladistics was built upon a central conept, that of the "monophyletic group", meaning the collection of organisms descended from the same common ancestor (population). The principles of cladism stated that each natural taxa should be monophyletic, therefore the research programme of cladism basically consists in uncovering monophyletic groups via the study of the pattern of descent (evolutionary tree). In this context, species turn to be segments of the evolutionary tree (pattern-wise) leading from one branching (node) of the tree to another, called "internodal species".

Common to all of these pattern-oriented concepts is that each is neutral with respect to the mechanisms outlined by the so-called process-concepts (BSC, ESC). Proponents of pattern concepts often argued that reproductive isolation, natural selection, adaptation etc. act in tandem and at variable rates in the formation and stability of species, but their joint effect is best recognized in the resulted phylogenetic patterns. So, what systematists should look for is the net result, instead of the causes. A hardly avoidable critique of this approach, and of pattern concepts in general, was based on a feature inherent to the relation of descent, being essential to this conceptual framework. The problem is that lineages of populations, the notion on which every pattern concept relies, are not trivially separable within the tree of descent, instead they are continuous in various ways: for one thing, lineages can be identified at various levels of aggregation, resulting in a nested

structure of taxa along the tree of life. A corollary of this observation is, for instance, that the CSC cannot guarantee, by its sole definition, that groups can be separated at the intended aggregation level of species. To put it differently, the issue at stake here is that of *taxonomic rank*, that is, the clear circumscription of the level in the taxonomic hierarchy of organisms that corresponds to biological, "real" species. In each case, pattern concepts are in need of a supplementary criterion to resolve this problem, i.e., to slice up the evolutionary tree at its appropriate joints. Indeed, proponents of pattern concepts do provide such rules, however, these rules are generally found external to the perspective of pattern-systematics. Rules of individuation often refer to geographical or ecological factors as well as reproductive isolation and intersterility, based on the context of application (taxa in question) – integrating process concepts into this alternative framework as well. Nevertheless, a clear methodological import of this line of the debate was demonstrating that any valid pattern concept should be based on two logical pillars. It should be supported by a so-called *grouping criterion*, as a necessary but, by itself, insufficient condition, usually naming the natural relation(s) on the grounds of which taxa are to be identified (interbreeding, isolation, descent, ecological similarity etc.). Furthermore, to make the list of conditions sufficient, in the case of lineage-based definitions it should also present a *ranking criterion*, that cuts the tree of life at the right points to arrive exactly at species taxa (the species category or rank). Up to this point of the story of systematics, however, ranking criterions seemed to inherit the issues of process concepts from which they were borrowed.

As a result of the flourishing dispute outlined above, in the course of which numerous variants of these focal concepts have been proposed, a rich inventory of candidate species concepts accumulated during, mainly, throughout the 80's and 90's. An outstanding review of these definitions from Mayden (1997) counted no less than 22 conceptually distinct definitions. The discourse, mostly located in the intersection of systematics

and evolutionary biology, can also be conceived as a research programme to identify the factors that are both (1) causally relevant and (2) conceptually sound to the species category, assuming that the latter cuts out real patterns in nature. A major tendency unfolded within this programme was the quest for a "common cause"-like key to species. This tendency is demonstrated by the abovedescribed attempts, whereby peer responses to challanged species definitions seem to follow a scheme known as "theory reduction" from the philosophy of science: new proposals ambition to generalize from previous concepts, both to incorporate them and to gain a higher level of universality (explaining phenomena that precursor concepts failed to explain). Looking at the path leading from the BSC to EvSC is a good example: the "own evolutionary tendencies" of taxa, which feature serves as the basis of delineation by the EvSC, was conjectured to be the net result of reproductive isolation (BSC) and other mechanisms, not being, however, reducible to any particular such process.

Despite all these historical efforts, the Species Problem has been, and is still well and alive. In the literature, one can find various theoretical attempts to resolve, assess, evaluate or explain the deep causes, durability and long-lasting nature of this issue. A striking feature of this background literature is that it goes far beyond biology, either theoretical or applied, in terms of contributing disciplines and scientific/scholarly fields. The Species Problem induced a type of debate that is rather peculiar in the modern history of science. This type is characterized with the actual interaction of distant scientific and scholarly specialities, namely, biology (life sciences) and philosophy (humanities), in the first place. Upon this feature, in order to cover the entire discourse, it is reasonable to coin the term *interdisciplinary species problem*.

The Interdisciplinary Species Problem (ISP)

The standard story above can be characterized as told from the perspective of biological systematics. However, the historiographer (and philosopher) of science would point towards an important aspect lacking from such a methodology-oriented reconstruction, which, however, plays a crucial role in the development of the problem. The late XX. century species problem involves and is deeply affected by a specific subfield of the philosophy of science, the philosophy of biology. The role of this scholarly field is, in part, a legacy of what we called the "classical species problem", which can be (and is often) reinterpreted as a metaphysical–ontological question: are species natural kinds? Do they exist, or are real/natural entities at al? Just as evolotinary biology, the issue of the ontological status of species has gained a radically new meaning due to the paradigm shift, which was induced by the modern Synthesis. In the sixties and the seventies the biologist Michael Ghiselin and the philosopher of science David Hull formulated the so-called "individuality thesis" (species–as–individuals; SAI) with respect to the ontological status of species, according to which species taxa are not kinds or classes, but, ontologically speaking, individuals (Ghiselin, 1974, Hull, 1978). Ghiselin labels this proposal a "radical solution to the species problem", indicating, basically, the launch of a new research programme. This programme should be aimed at formulating a definition of the species concept, that, beyond being methodologically valid, is compatible with this ontological claim. The solution lies exactly in this criterion: a true and acceptable (scientific) definition of the species category can be selected – so the argument goes – from the inventory of species concepts outlined above, if such a constraint is taken into account. The critical reception of this thesis induced a broad and still-present controversy among and between biologist and biophilosophers. The controversy has resulted in a huge variety of (biological, ontological, logical, semantic, etc.) arguments for and against the thesis on one hand, and criticism of the programme to apply the thesis as a new paradigm that

diminishes the problem in a natural way (Ghiselin, 1981a). In addition, a wide range of now-canonic issues have emerged in biophilosophy addressing the dissolution of the SP, focusing on the abundance of species concepts, the reality and objective nature of the Species Category (pluralism – monism, realism – antirealism, etc.)

The basic enterprise of this book is to contrast the above story with empirical evidence, through the application of a rich and novel theoretical framework called science mapping.

Chapter 2: Data collection on the Species Problem, 1975–2010

Identifying sources of bibliometric data

The empirical reconstuction of the modern species problem is based on the "tomography" of bibliographic and citation databases. In order to build a standardized, machine-readable corpus suited to complex science mapping purposes, candidate databases should meet the following minimal set of criteria:

- coverage is international,
- coverage is multidisciplinary, that is, documents (journals) are being indexed with respect to all scientific/scholarly disciplines and fields,
- coverage can be characterized by a historical depth that is in accord with the timeframe under study (which is, in our case, the late decades of the XX. Century)
- references of, and citations to documents are available to analysis and it is feasible to process them in a computer-aided fashion (this criterion is inevitable if one were to build the history of the flow of ideas in the discourse),
- data are – at least with institutional academic licences – "publically" available.

The above criteria, at the time of the underlying study, is mostly met by the famous Web of Science service (established by the Institute for Scientific Information – ISI – and commercialized by Thomson–Reuters, herefater referred to as WoS) , which is, in fact, a collection of citation databases. In order to cover a representative corpus, three WoS databases were selected for obtaining data, that jointly coverded the landscape of the modern science system:

- The Science Citation Index (SCI),
- The Social Science Citation Index (SSCI), and
- The Arts and Humanities Citation Index (A&HCI).

Such a choice, concerning the data sources of our study, raises a question of a "doctype-bias". It is well known that it is mainly journal articles that are covered by the Web of Science. Also, one may object that the historical coverage of data source would be insufficient for our purposes, since the WoS provides records dating back "only" to 1975. However, on a closer look, both features fit well to our research goal (or, at least, compatible with it). To the first issue, we should point to the main feature of our study to rely on citation relations, that is, on references (and not only on source documents) as well. By extracting references from WoS records, the documents not represented as a source record also come to light. As to the time window, the nature of the discourse provides a comforting answer: the modern story of the species problem steered up in the late decades of the XX. century, producing most of its corpus as well. In fact, what requires an empirical clarification in its history tends to be concentrated within these very decades being covered in WoS. This was the reason behind setting the limits of the publication window between 1975 and 2010 (again, only for source publications but not for the works cited by them), a period that safely covers those dynamic knowledge interactions in the discourse that constitute our interest.

Constructing the core

In the first and fundamental step towards building the corpus, a so-called core corpus was identified, with the corresponding bibliographic and citation data being harvested from the SCI, the SSCI and the A&HCI. By "core" we meant a collection of documents that could be directly assigned to the discourse as contianing explicit referrals to the species problem. The respective database query was therefore defined to include all the records each containing any of the following terms within its title/set of keywords/

abstract: “species problem”, “species definition”, “species concept”. As can be seen, data retrieval in this initial step was purely topic-based, as an attempt to avoid the potential exclusion of relevant works from the corpus, since such a query did not put any constraints on the set of fields, journals, authors etc. entering the sample. The resulting corpus included N=1605 documents for the period 1975–2011.

Expanding the corpus along the cited works

In an attempt to gain a comprehensive historical coverage on the topic, we have extended the collection of core documents via an iterative analysis of aggregated references included in each step. The rationale behind the procedure described below was to reveal the “latent” part of the corpus, that is, the body of literature recognizable only through citation relations (not being explicitely linked, by textual descriptors, to the species problem). This approach was based on the assumption that analyzing references in an iterative fashion would lead to a reasonably saturated (or complete) set of documents that is relevant for the species problem, and which also sets up a citation network along which the flow of ideas could be captured, uncovering the historical development of the discourse. The iterative method consisted in the following stages:

(i) In the first step of the process, references from the initial corpus were processed, and the corresponding set of source documents was obtained from the WoS databases. This additional publication record was then added to the pool of already collected papers (while duplicates were filtered out).

(ii) The above procedure was re-iterated for the extended set of source documents.

We repeated this method in further iterations, until reaching a collection being fairly "closed" under the citing relation, that is, a collection that contained all the – topic-relevant – papers referred in the discourse. To assure such a convergence, references were filtered by a threshold imposed on their frequency: papers cited above this threshold were, in each round, considered relevant for the topic. The threshold value was increased (non-linearly) for each iteration, based on the assumption that the farther we get, along a series of references, from the core set of papers (in terms of corpus generations), the less related references will be to the topic.

Interestingly, with this setting, the procedure converged in the third generation of papers, indicating that almost all relevant references were present after two iterations. Finally, for the discourse of the species problem, we arrived at a final record of approximately 5700 papers (the main statistics of the procedure are summarized in Table 0 below.) In particular

The 1. generation, that is, the core corpus, that consisted of n=1605 docs, contained a total of n ≈ 51 thousand unique references. The threshold for references to be considered relevant was chosen at a level of 3 occurences throughout the whole sample (f=3). Based on that criterion, a collection of n=3200 cited docs have been obtained from WoS, as the next (or, in fact, ascendent), second generation of the corpus.

The 2. generation. Proceeding with the references within this second round of papers (after filtering out those documents that were alrady cited by the 1. generation core), a broad set of 62 thousand uniqe references were identified. However, in order to draw on a subset relevant to our study, the the threshold for reference selection was significantly increased: the constraint on "eligible" references was to exceed an occurrence frequency of ten (that is, docs cited at least 10 times throughout the extended corpus were attached to the sample). Such a relatively stong relevance criterion (as compared to the previous one) can be motivated by the assumption

that broadening the corpus along generations of references imports distinct, but overlapping bodies of literature, through related research fields, methods, etc, tackled by the species debate. A further argument might be the unbalanced bibliometric behavior of constituent fields: citation density considerably differs both between and within relevant disciplines, as can be illustrated by contrasting the life sciences and the philosophy of science (as between disciplines), or genetics/molecular biology versus ecology (as within a broad discipline). In both cases, a paper of equal relevance tends to be cited more often in the former than in the latter, resulting in an overrepresentation of life sciences or genetics, and suppressing contributions from philosophy or ecology (respectively). The practical choice of f=10 seemed to provide a reasonable tradeoff to keep the cohesion of the corpus, in terms of being representative of the Species Problem as such (instead of some overlapping discourses with different foci), while not excluding relevant documents due to different cultures of scientific communication. Along these lines, a further set of n=850 docs have been identified and amended to the existing corpus.

The 3. generation. Exploring the references included in the second generation described above led to the 3. generation of references. In this particular step, the threshold value was kept unchanged (not being increased). This choice was confirmed by the observation that even with this value (*f*=10), only two (!) further documents could enter the corpus, that is, appeared to be relevant to the discourse. With such a definite indication of arriving at a practically closed (or, "complete") set of papers representing the history of the Species Problem, the construction of the corpus was considered to be finished.

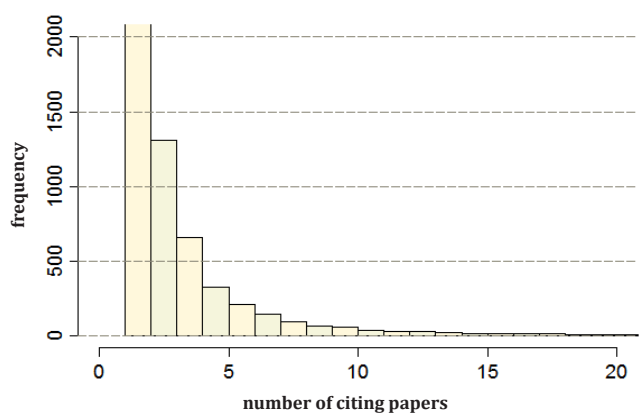
In sum, the iterative procedure for building our large-scale bibliographic sample has proven to be convergent., along with the strategy of strenghtening the constraints on eligible reference in a step-by-step fashion. As a direct result, the size of the collection after all the iterations

amounted to n=5679, which, after filtering out duplicates or poorly identifyable references, left us with a corpus made up of approximately five thousand papers (n=5173).

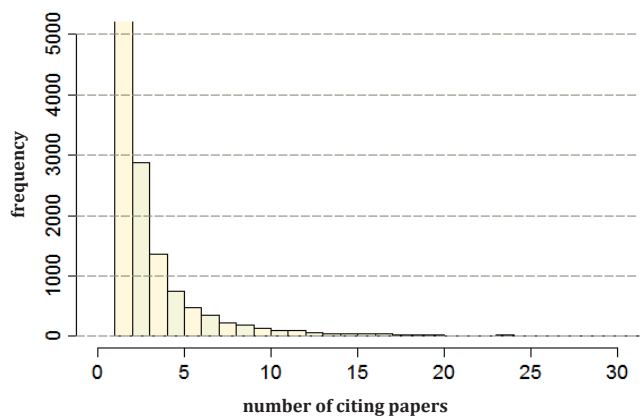
Table 0. Statistics of iterative corpus collection on the Species Problem based on WoS databases

Iteration	*No. of source documents*	*No. of references*	*No. of unique references*	*Threshold value*	*No. of relevant references (retrieveable)*
Initial corpus	1605	93 943	50 668	3	3223
2. generation	3223	155 742	62 574	10	851
3. generation	851	14 991	5305	10	2
Total	**5679**				

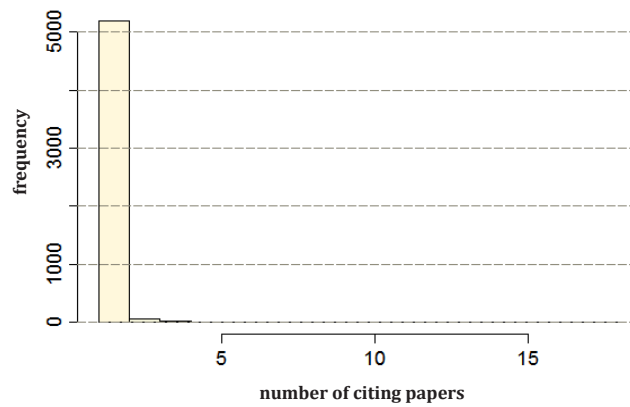
Frequency distribution of the appearance of unique references within the core corpus (1st generation)



Frequency distribution of the appearance of unique references within the 2nd generation



Frequency distribution of the appearance of unique references within the 3rd genration



Chapter 3: Delineating and modelling the discourse as a citation network

In order to prepare and facilitate knowledge discovery along this large-scale longitudinal bibliography, we have organized the final corpus into a citation network. The large directed graph obtained from the document set consisted of all papers included in the full corpus as its nodes; edges represented the (direct) citing relation between any two documents.

Modelling the corpus by citation relations was motivated by, at least, two conceptually different, but interrelated reasons. On one hand, the fundamental goal of our work is to reveal the large-scale causal–historical structure of the species debate, the best bibliometric indicator of which is one document (author etc.) citing another. Consequently, the pattern(s) we are looking for is (are) best detectable by transforming the corpus into a citation network. On the other hand, the corpus collected by the method above is, expectedly, much broader than the specific debate on species. The explanation is, implicitly, also outlined above: navigating through references by an iterative method imports a wide range of overlapping subjects, which all contribute to, but jointly underspecify the particular subject under study. However, it is the very citation network set up by collected documents, and, even more precisely, the topology of this network, the uncovers the boundries of the discourse.

In particular, as documents belonging to the discourse can be assumed to affect each other either directly or indirectly, one could expect the species debate to show up as a more-or-less separable or coherent part (subgraph) within the citation network. In terms of network analysis, the Species Problem is likely to form a community, or – assuming a higher degree of separation – a connected component of the network (whereby each document is connected, direclty or indirectly, to all other members of the component, but unconnected to others outside it). Based on these considerations, a second analysis has been conducted on the – now re-modelled – WoS-corpus, in order

to identify the proper sample for representing the history of the problem. This second analysis consisted of identifying the coherent subgraphs of the citation network via network analysis tools, and selecting the one (or those) corresponding to our research question.

The analysis of the whole citation network revealed that the large-scale structure incorporates two isolated part, namely, connected components. In other words, two sets of documents were identified: within each of these sets, there existed a citation pathway connecting any of its member to another. At the same time, no pathway could be found between those sets, suggesting that what we uncovered were two isolated subdiscoourses related to the species concept. One of these components is the so-called "giant component", covering most of the papers in the network, its size being n = 4382. The other was a significantly smaller citation graph (with orders of magnitude), counting n=53 papers.

In order to assess the correspondence between our subject and the identified citation graphs, a further approach from network analysis has been utilized. The basic idea was to pinpoint the most "relevant" works in both networks: whereby the titles that highly affect all other documents through the respective citation network are the ones centered around the species definition (as such), we can consider the corresponding component as the citation graph of the species problem. Relevance for a document, in this case, can be equated with a specific position in the citation graph. In social network analysis (SNA), well-known measures of relevance in a communication network have been established, called measures of centrality. *Betweenness centrality* is a proxy to the importance of a paper in organizing the whole graph (indicating the extent to which the paper mediates between other papers in the flow of knowledge characterized by network topology).

By applying this measure, we calculated betweenness centrality for each document (node) in both the giant and the small component of the entire citation graph, in order to obtain a proxy to the position-based relevance

of papers within their own discourse. We, then, imposed a centrality-based ranking of documents on both subnetworks, obtaining, therefore, the works with the highest relevance in the above sense. The results of this exercise clearly showed, that the giant component can be identified with our subject, the polemy on the species concept, since the first (most highly ranked) documents in the ranking turned out to be seminal works well-known for their contribution to the debate. On the other hand, the small component also exhibited a clear identity, as it covered the discourse on the mathematical modelling ecosystems, whereby the "species problem" refers to a different research question, namely, the role of species as building blocks in ecological systems, like foodwebs (as reflected in terms such as the "keystone species problem"). Therefore, for the purposes of our research program, we isolated the content of the giant component as our final sample to discover the organization of the species debate. The results of this process, along with the main quantitative characteristic of the selected fragment of the corpus (size) are summarized in the table below. (The seminal titles referred to above, the full content of the components along with the centrality values and the relevance-based rankings are available in the form of a database supplementing this book).

Component of the citation network	Size (# docs)	Scientific/scholarlydiscourse Represented
1	4382	Debate on the species concept
2	53	Modelling ecosystems

Chapter 4: Scientometric methods of mapping science

In this chapter we attempt to provide a concise overview on paradigmatic science mapping methods. Our aim here is to outline the existing inventory of this knowledge representation methodology in order to lay the groundwork for the main research project: the advancement, elaboration and integration of the selected methods for our purposes. To this end, the presentation of SM will cover the elementary methods through organizing them into a taxonomy which we can use for further reference when deloping our toolkit (out of the elements of this inventory). Since all the elementary methods discussed below employ network models, we categorize them under the applications of network analysis methods.

Network methodologies for uncovering the organization of scientific aggregates are best arranged along two dimensions at the operational level: (1) the type of bibliometric indicator(s) used, and (2) the type of network model constructed based on the selected indicator(s), along with the represented construct regarding the dimensions of scientific knowledge. Main categories of models and their corresponding applications are the following.

- Methods of reference pattern mapping

The most paradigmatic methodology in the bibliometric mapping of the structure of scientific fields is based on the indicator set provided by the references of papers. The main assumption behind this is that references jointly constitute the intellectual background or the "knowledge base" of papers: therefore, analyzing the aggregated reference set of a corpus representing a field at any given time slice uncovers its cognitive structure. References provide multiple bibliometric indicators for analysis, including

- cited documents (D),
- cited authors (A),
- cited sources (typically journals) (S),

each conveying a different aspect of the field structure (see below), and posing different requirements on, or challenges for computing power. Methods utilizing the above indicators fall into two major categories.

Bibliographic coupling (BC). Source documents representing a field are clustered based on their degree of sharing the same references (in terms of documents, authors or sources). Various measures of similarity and clustering techniques are used (Kessler, 1963).

Co-citation analysis (CC). References in source documents are clustered based on their frequency of being co-cited by the source document set (in terms of full references, included authors or sources). Again, various measures of similarity and clustering techniques are used. (Small, 1973)

The two basic techniques are the converse of each other: in BC, source documents are grouped via references; in CC references are grouped via citing source documents. Prototypic approaches addressing the organization of science using reference-based mapping,are the following.

Intellectual structure of fields. A rather traditional approach, author-co citation analysis (ACA) is often used to detect and visualize the cognitive structure of research fields. ACA is the combination of (A) and (CC), as it takes cited authors as the unit of analysis, and yields author clusters based on their "co-citedness". Clusters are conceptualized as research communities concentrated around a specific research topic, thereby mirroring the thematic composition of the underlying field.

Disciplinary organization of science. Variants of source co-citation analysis (SCA) are used in large-scale approaches addressing the global structure of science. For constructing a "global science map", several researchers used SCA. In models of Moya-Anegon *et al.* (2004), specialties of science represented by Subject Categories in the ISI databases are subjected to analysis. Subject Categories are source indicators, as they are introduced to categorize journals in the database. Using source documents, and substituting cited sources for the corresponding Category, the proximity of Subject Categories is calculated measuring the degree of their co-citedness throughout the whole corpus. As a result, a proximity network of Categories (specialties) is obtained, which can represent the global structure of science.

The method described above is thus a combination of (S) and (CC). A somewhat different approach has been introduced by Leydesdorff and Rafols (2009), whereby Subject Categories are related upon their citation patterns, namely, by their degree of co-citing the same Subject Categories, resulting also in a proximity network. Consequently, while the maps of Moya-Anegon *et al*. belong to the class of co-citation analyses, the latter approach is an example of bibliographic coupling on (aggregated) sources, i.e. a combination of (S) and (BC).

Global map of scientific paradigms. The most detailed picturing of the scientific landscape to date has been achieved by the "paradigm mapping method" (Boyack et al 2005, Boyack 2009). A paradigm in this setting is operationalized as a frequently co-cited group of references, reflecting a cohesive topic or specific subject of research. The global paradigm map of (Boyack et al. 2005) has been generated by processing the content of the Scopus database: the full references of source documents were subjected to co-citation analysis, and clustered based on the resulting proximity matrix. The procedure yielded something that may be considered the global map of scientific paradigms at a given time slice, unraveling a cluster structure

at an extremely high level of granularity. Since the method relies on full references, it qualifies as an instance of document co-citation analysis (DCA), hence combining (D) and (CC). It should be noted that, due to the outstanding amount of documents and references, the method requires considerable computing power.

- Citation-flow mapping

By utilizing the nature of scholarly citation, corresponding indicators naturally enable science mapping to empirically address dynamic or historic aspects of science. To detect and visualize the flow of information, the spread and transformation of ideas, or the development of conceptual systems, citation-flow mapping is utilized, sometimes called "algorithmic historiography" (Garfield et al. 2002). The method implies the construction of a citation network of papers either over a given timescale or about a given topic. This network is conceptualized as representing the patterns of information flow. The network, in this case, is static in the graph-theoretic sense, but represents longitudinal, i.e. dynamic content, mapping a process along its time dimension. A genuine implementation of the concept can be found in *HistCite*, a software providing citation flow analysis based on the ISI databases.

- Author-based mapping

Another common bibliometric indicator for science mapping is authorship, or, rather, co-authorship. Many studies have attempted to reveal the composition and development of research fields by analyzing the author-network encoded in the corresponding publication corpus. The collaboration of authors resulting in joint publications is conceived as a shared research interest, upon which the "visible colleges" of a field or a discipline can be identified. Technically, the analysis of co-authorship patterns proceeds by

first extracting the network of authors from a bibliographic dataset, where ties stand for two actors co-authoring at least one paper. This network is then subjected to community detection methods, and decomposed into coherent author clusters, that is, into scientific communities. Though this approach is, at face value, just an application of social network analysis, and so targeted at the social organization of science, the factors behind group formation, such as working on close topics, make it capable to grasp cognitive organization as well.

Co-author networks are often studied from within the network science perspective, irrespective of their use for science mapping purposes. Various generalizations have been made on the structure and dynamics of such networks, drawn from assigning them to the class of scale-free networks. For instance, the growth of co-author networks by "preferential attachment", a process responsible for many scale-free structures, is also a well-known claim (Barabasi et al. 2002). Other studies more directly in the SM domain have investigated general co-authorship patterns and -dynamics in relation to the evolution of research fields (Bettencourt et al. 2009). The report below will heavily utilize the results of this latter approach.

- Mapping conceptual structures

A different, and frequently utilized set of bibliometric indicators is constituted by the textual descriptors of documents. Descriptors in this category include keywords associated with documents, title words, or the characteristic words obtained by text mining from either the abstract or the full text of papers. As can be seen from the list, the methodology may involve natural language processing and text mining procedures, which makes this approach relatively expensive compared to the utilization of directly accessible metadata types. For the sake of simplicity, we describe the methodology below using the case of author keyword analysis. Author keywords are concepts chosen by the

author to jointly convey the content of the respective paper, being readily available in many scholarly databases among the metadata of documents. Therefore, processing author keywords does not require text mining or other linguistic pre-processing.

Since keywords are meant to provide immediate access to the content of papers, their association patterns in large-scale document sets are considered as (1) the most directly interpretable and (2) the most fine-grained mapping of the cognitive structure of the underlying field. The method is referred to as co-word analysis: first, a pairwise association of keywords is measured in a document set, based on the frequency of their co-occurring in documents. Next, this association matrix is decomposed, either by direct clustering or (conceived as a proximity network of concepts) by community detection methods yielding groups of closely related words. These groups are then interpreted as thematic clusters comprising the field under study.

Co-word analysis, the alternative method for building a representation of the cognitive structure of science, is often contrasted with co-citation analysis as being suited to somewhat different tasks of science mapping. The main argument is based on the recognition that references encode the past, or background of a paper, while keywords are "of the same age" as the source document itself. Hence, co-word analysis is argued to be more capable of grasping ongoing trends or emergent topics than co-citation patterns, which may not react to rapid changes or to the appearance of genuinely new directions (cf. Chen 2003). Sensitive as it is, the co-word approach has also been challenged by theoreticians such as Leydesdorff (1997) who pointed out that, among other things, the association of words without a sufficient information about the embedding context leads to uncertain interpretations and risky semantics, violating the validity of evaluating these maps.

To achieve higher levels of accuracy and expressive power, the basic methods summarized above are also often combined, resulting in various hybrid methods. (cf. Janssens 2008; Glänzel & Thijs 2017).

Chapter 5: Methods for mapping knowledge flow via citation-analytic models

As a methodological advancement of existing tools in science mapping, in the present chapter we propose three novel bibliometric methods for the discovery of the Species Debate (and, in fact, for any debate in the history of science). The common feature of these methods is that each relies on the citation network explored within the corpus. As such, these methods aim at mapping the cognitive connections induced by the actual knowledge flow between documents, providing a valuable means for reconstructing the causal processes organizing the conceptual – thematic structure of the discourse.

Model 1.1: Age-sensitive bibliographic coupling

In science mapping, bibliographic coupling (BC) has been a standard tool for discovering the cognitive structure of research areas, such as constituent subareas, directions, schools of thought, or paradigms. Modelled as a set of documents, research areas are often sorted into document clusters via BC representing a thematic unit each. In this chapter we propose an alternative method called age-sensitive bibliographic coupling: the aim is to enable the standard method to produce historically valid thematic units, that is, to yield document clusters that represent the historical development of the thematic structure of the subject as well. As such, the method is expected to be especially beneficial for investigations on science dynamics and the history of science.

Bibliographic coupling (BC) is a long-established method in science mapping. Its main aim is to detect, within a set of publications, groups or clusters that share a common intellectual background, and, therefore, can be conceived as each representing a particular research problem, program, approach or school, depending on the interpretation. To this effect, the method relies on

references, usually conceptualized as conveying the intellectual background of the corresponding papers. The basic principle is that the relatedness of any two papers is a function of the number of references they have in common.

Since the introduction of the method within bibliometrics (Kessler 1963, also cf. Small 1973), the method of BC has been effectively applied in many contexts, basically in its original form. In this chapter we propose a refinement of BC that takes into account a further parameter of common references: beyond their (usually normalized) number it also incorporates the (respective) age of them. We call this method *age-sensitive bibliographic coupling*. The reason for and our expectations on this alternative method is best communicated with the help of an analogy from biological systematics.

A striking similarity between reference-based science mapping and evolutionary biosystematics is that both attempts to detect groups of related actors based on common ancestors. In the case of science mapping, biological descendancy is to be replaced by citation links, or "intellectual descendancy": a reference can be viewed as an ancestor of the citing document. However, as a disanalogy, biosystematics defines the degree of relatedness as conditional on the "age" of common ancestors: on the evolutionary timescale, the more ancient their common ancestor is, the less related two species are, while the more recently they originated from a common predecessor, the closer they stand in systematics. As a result, biosystematics is capable of setting up a categorization where groups also reflect the history of their formation.

We claim that these considerations can be adopted for bibliographic coupling as well to gain similar advantages. Our modified basic principle of BC, therefore, would formulate in the following way: the more recent references any two papers have in common, the higher the degree of their relatedness is. That is, the (intellectual or cognitive) relatedness of any two papers is a function of the (1) number and the (2) age of references they have in common.

Addressing the age of references in bibliometrics is, by far, not a new idea, – consider, for example, the classical Price index (Price 1970), conveying the age distribution of the intellectual background – nor is the assumption that the subset of references published more recently is indicative of the particular direction of research a paper belongs to, as contrasted to "older" references, characterizing the broader thematic context. However, approaches linking these observations to bibliographic coupling have been rather rare. One such example is the study of (van Raan 2005), addressing the behaviour of BC. For a sample of documents to be structured by the method, Raan partitioned the set of aggregated references into two age groups based on two consecutive time windows, producing a cohort of "old" references and another of "young" references. The application of BC on sample documents using the old cohort and the young cohort, respectively, resulted in similarity networks within the sample with different structural characteristics (degree distribution). Based on these results, Raan argued that the young cohort, that is, recent references, is better suited to classify documents according to their intellectual relatedness, which is in accord with the assumption on the role of immediate cognitive ancestors.

As contrasted to this latter approach, our goal is not to filter the set of references so that an improved precision of clustering could be achieved via BC, reflecting exclusively the closest and most timely relations. Instead, we aim at the "whole picture", inside which all relations are made visible, but still (historically) distinguishable: relying on the entire, unfiltered (and aggregated) list of referred works, we intend to incorporate age as a factor into the method, and potentially obtain clusters being differentiated in this respect: some reflecting a closer, some looser internal historical relatedness. The rationale behind is the same as in the case of biosystematics: by age-sensitive bibliographic coupling we expect to map a research area not only in terms of "thematic directions", but by revealing real, historically (causally) connected parts of the discourse.

Since (1) we were primarily interested in the period where the debate became most intense and accelerated (so that empirical methods are helpful to clarify its structure), and (2) the selected data was also required to "contain enough references", potentially reaching back to all historical layers of the debate, we took a smaller time window for our analysis. As confirmed by the distribution of the core corpus over publication years (Figure 1), a period starting from the '90s was meeting the intensity criterion, and was late enough to reflect existing directions. We took a fraction of the whole corpus accordingly, covering a decade being a "burst" in the dispute. This fraction contained about 400 records. We pruned it by eliminating those few that did not share any references with the rest (not being related to the problem, in this sense). Our final sample, therefore, contained the fragment of the base corpus published between 1990–2000, with N=386 papers.

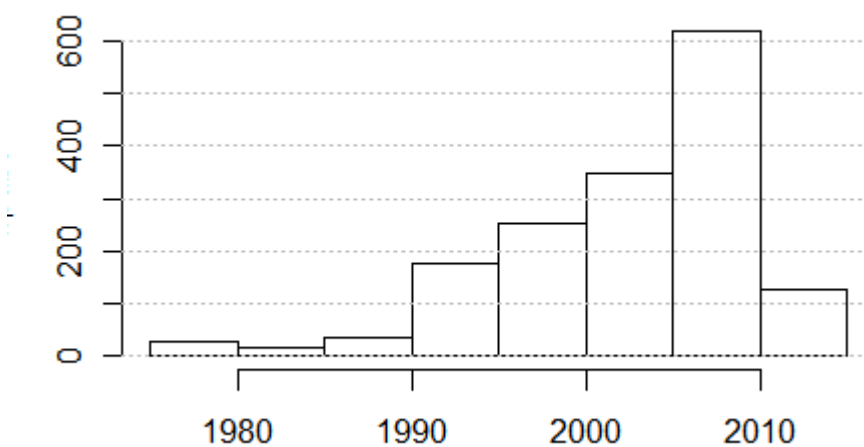


Figure 1. Distribution of the corpus collected on the Species Problem over publication years (before sampling and pruning). On the y-axis, labels stand for absolute frequencies.

The altered method of bibliographic coupling

Bibliographic coupling of a set of publications, in the classical case, is based on the number of their common references, in a pairwise manner. This relation can be conceived as some sort of similarity (or distance) between the two vectors of references of any publications *p1* and *p2*, respectively. These vectors are usually represented as dichotom sequences with values {1,0}, denoting the presence/absence of a publication in the reference set (this method also implies that such vectors are built over the aggregation of the references of each pub in question to make them comparable – practically, these are the rows of a publication-reference incidence matrix).

More formally, the method of determining the relatedness of pubs within classical bibliographic coupling may be presented as follows. Given any two publications P_1 and P_2, consider the vectors REF_1 and REF_2 of their respective sets of references. These vectors are best conceived as of length n, where n is the number of all references belonging to either P_1 or P_2. Based on these same tuple of referred publications, $REF_1(i)$ denotes whether the i-th reference is present among the references of P_1, and may take the corresponding value of 1 or 0 (the same goes for P_2). In this setting, the basic similarity between the two publications is given by

$$S_{\boldsymbol{B}}\ (P_1, P_2) := \sum_{i=1}^{n} REF_1(i) \times REF_2(i).$$

In verbal terms, $S_{BC}(P_1, P_2)$ is the absolute number of references shared among P_1 and P_2. This amount is usually subject to a normalization procedure accounting for the size of the reference sets of P_1 and of P_2, respectively, for it is often argued that having the same amount in common out of an extensive background (of which the shared part is a relatively small fraction) makes pubs less related, than if this same amount is a substantial part of the references for any member of the pair. In our study, however, we used

this measure in its raw, non-normalized version, mainly for the reasons of comparison with our age-sensitive indicator (see below).

In order to implement the idea of age-sensitive bibliometric coupling, we altered the abovedescribed method of BC in two steps. The procedure was based on the publication–reference incidence matrix constructed from publications in our material.

Step 1: Weighting

At first, an indicator of the age of references has been introduced. To systematically account for this feature of reference publications, each component of the presence/absence vectors was weighted according to the publication year of the corresponding reference. This procedure yielded a weighted reference vector for each source publication:

$$REF^{W}(i) := REF(i) \times f(Pubyear(i)),$$

whereby $REF^{W}(i)$ is the weighted value of the i-th reference within the vector of references $REF(i)$, and this weight is given by a function of the publication year of the i-th reference, i.e. $Pubyear(i)$.

Practically, this kind of modification of a presence/absence vector replaces the value "1" of each reference of the source publication with a time-dependent weight, determined by the weighting scheme. In order to reflect our "phylogenetic" notion of relatedness, we defined the particular weighting scheme (the function $f(Pubyear(i))$ in the formula) according to the following criteria:

(1) The more recent a shared reference is, the closer relatedness of source documents it should represent.

(2) Classical topic-related literature should reflect distant kinship when referred by pubs, while shared recent literature reflect close kinship. Furthermore, as we intend to amplify the effect of having classical vs. recent common ancestors in drawing relatedness (so that recent kinship and more ancient kinship could be separated), it is assumed that differences between the age of classical (old) publications contribute less to relatedness, than age differences in the recent literature.

In the scheme chosen for weighting, criterion (1) is realized by weights being defined as increasing by publication years. This procedure assures that, when subjected to the similarity measure introduced below, recent references contribute more to document similarity than older ones. Criterion (2) is met by rewarding a reference for being timely, via determining weights as a non-linear function of time (publication years). In particular, we used an exponential function of the rescaled years of publication, the parameters of which were experimentally set to enable the scheme conveying the age effect of the intellectual background, in the case of the topic under study:

$$w(Pubyear) := \mathrm{scale}_1\left(\beta^{\ \mathrm{scale}_2(Pubyear)}\right),$$

whereby *Pubyear* is a year of publication (age), *w(.)* is the associated weight, and $\mathrm{scale}_2(Pubyear)$ designates a linear rescaling of the series of publication years within the interval [1,10]. The immediate result was also rescaled within the interval [1,100], indicated by $\mathrm{scale}_1(.)$, to produce intuitive weighting scores for references. Figure 2 graphs the weights associated with years of publication. It can be observed that (due to the distant origins mentioned above) references to the ancient – e.g. medieval or XIX. century – history of the problem, ranging from the XVI. century to the beginning of the XX. century are almost equally weighted, their contribution being kept at a low level. The weighting is becoming rather progressive from the 1960s, and the slope of the curve increases by roughly twenty years (at the beginning of

the ’80s, and that of the second millennium). This scheme is in accord with our aim to detect the accelerated development of the topic in the XX. century, and also with descriptive studies characterizing similar periods of problem development along the timescale.

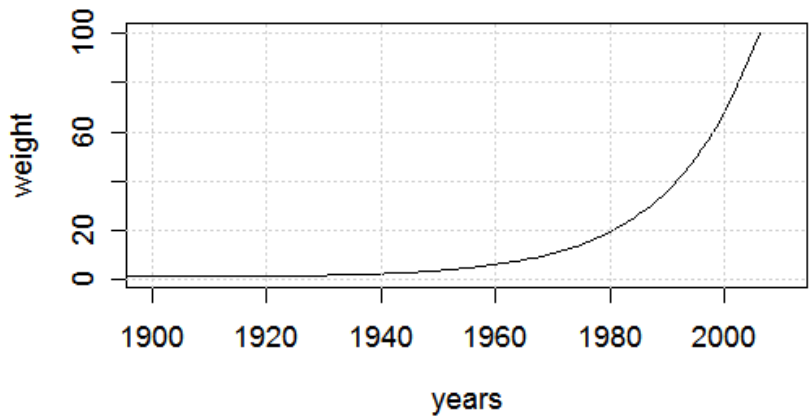


Figure 2. Weights associated with publication years according to the weighting scheme used for the age-based ranking of references

Step 2: Similarity measure

As the second step of the method, the degree of relatedness of source documents was calculated based on their weighted reference vector. In particular, we applied the basic similarity measure S_{BC} of bibliographic coupling explicated above, to each pair of such vectors obtained for source documents in the sample. This resulted in a measure

$$S_{B}^{W}\left(P_1, P_2\right) := \sum_{i=1}^{n} REF_1^W(i) \times REF_2^W(i),$$

Where $S_{B}^{W}\left(P_1, P_2\right)$ is the weighted (or age-sensitive) similarity of publications P_1 and P_2, while REF_1^W and REF_2^W are the weighted reference vectors belonging to P_1 and P_2 respectively.

In practice, according to this measure, the more recent references are being shared by any two publications, the more closely related (similar) these pubs will be. Defined in such a way, this indicator is not normalized (e.g. doesn't control for the number of references that the two publications contain, separately), but since we are interested in the effect of age (weighting) of shared references, as disentangled from any other effect , we used the measure as such: this choice allowed us to contrast the results directly with the core of the classical (unnormalized) approach, whereby the same common references are counted but not age-weighted. More importantly, by this definition we obtain a fine-grained relation between publications, even analytically. Consider a publication *P1* that has the same number of common references both with *P2* and *P3*, but with recent publications shared with the former, and with old publications shared with the latter. On the classical account, *P1* is equally similar to *P2* and *P3* (since only the amount of shared references matters). However, on the present account, *P1* is much more similar to *P2* than to *P3*, due to the contribution of recent background literature to the similarity value.

Clustering of source publications

Though not specific to the altered procedure of bibliographic coupling discussed so far, a still relevant step of the method is the actual "coupling" (or grouping) of publications, that is, the clustering based on the weighted similarity matrix. For this purpose, a type of hierarchical clustering was selected, and imposed on the distance matrix obtained from the original similarity matrix. We applied the average clustering method, as the resulting hierarchy turned out to be, among those produced by other available methods, best fitted to document distances. (This latter fit was measured by the so-called cophenetic correlation, and yielded a value $cpc = 0.7$)

In order to detect the cluster structure at a fine-grained level, we avoided to cut this cluster tree at a predefined height, as such a trade-off would have resulted in overlooking groups with varying "internal cohesion". Instead, an approach called *dynamic cutting* was utilized, as developed and detailed in (Langfelder–Zhang–Horvath 2008). The main advantage of dynamic cutting compared to the traditional cutting-at-a-specific-level approach is the sensitivity to the shape of the dendogram and to nested groups. Due to our phylogenetic view on BC whereby closer and looser relatedness is assumed to be definitive of groups, we expected nested clusters (that is, groups to be recognized at different levels of cohesion). Therefore, this tool seemed to suit our needs quite well.

Having defined age-sensitive bibliographic coupling (*asBC*) on the basis of the classical approach (*cBC*), we subjected our corpus collected on the history of the species problem to a dual analysis. For the purposes of comparison, we applied both the classical, and the new method to reveal its cognitive structure. In what follows, the results of the two clustering exercises are presented and compared.

Quantitative comparison

To the effect of a first diagnosis to see whether the results of *cBC* and *asBC* could be expected to show a different picture of the corpus, the degree of similarity between the two groupings were estimated. We used two indicators thereof, (1) the Jaccard index plus (2) the correlation of cophenetic distances within the respective clusterings. The Jaccard index, in this case, could be interpreted as the relative extent of overlap between the two clusterings with a range of values [0,1], and yielded a value of $J = 0.3$, reporting a relatively small portion of document pairs that are judged similarly by both methods. Indicator (2) goes beyond this level of granularity, as it measures the change of relative positions each document has in the cluster tree based on *cBC*, when

recalculated via *asBC*. The correlation obtained was $r = 0.66$, indicating that the distances of documents within the cluster tree has moderately changed due to the age-sensitive grouping, that is, groups of documents are more closely or loosely connected on the new account (within in the hierarchical cluster tree). This observation is in accord with our expectations outlined in the previous section. In sum, the two diagnostics suggested that the age-sensitive version of BC generated a refined cognitive structure with different clusters, resulting mainly from the redefinition of document similarity increased or decreased as a function of the age distribution of references.

In more detail, the classical procedure, *cBC* resulted in a corpus divided into N=4 clusters, while the age-sensitive version, *asBC* yielded N=6 clusters. These numbers, already at this quite general level, suggest that *asBC* did result in a refinement of the clusters from *cBC.* This assumption is further corroborated by the size of these groups (that can be read off from Table 1, see below). While in the original case (cBC), 63% of the sample documents formed a single category, the age-sensitive version produced a more even, less uniform distribution with the first two groups accounting for 36% and 30% of the corpus, respectively. The remaining asBC-clusters were also in a par with the remaining cBC-clusters, that is, no degradation of group size according to the refined method could be observed (indicating small, less “proper” groups, outliers etc.).

cBC / asBC	**1**	**2**	**3**	**4**	**5**	**6**	*Sum*	*(×100) %*
1	89	107	0	27	22	0	*245*	*0.63*
2	0	0	28	0	0	0	*28*	*0.07*
3	17	0	0	0	4	0	*21*	*0.05*
4	34	9	5	4	1	39	*92*	*0.24*
Sum	*140*	*116*	*33*	*31*	*27*	*39*	*386*	*1.00*
(×100) %	*0.36*	*0.30*	*0.09*	*0.08*	*0.07*	*0.10*	*1.00*	

Table 1. Comparison of the clusterings obtained by cBC *vs.* asBC *via a confusion matrix.*

To put it another way, the new method seemed to split the largest (and, as unifying most documents, supposedly somewhat meaningless or hardly interpretable) cluster into smaller ones, that are expected to be historically more coherent (see the qualitative section below). Indeed, the so-called confusion matrix of the two groupings has the same implication (Table 1.). The confusion matrix is a cross-table of the two clusterings, reporting the joint distribution of sample documents within both sets of clusters (so that the relation of *cBC*- and *asBC*-groups could be examined). The rows of Table 1 correspond to the four clusters drawn via the cBC-method, as the columns to the six new clusters from the asBC-method.

As is apparent in the matrix, most affected by the re-partitioning of the species problem literature is the *cBC*-cluster no. 1, that has been split into mainly two, similar-sized groups, *asBC*-clusters no. 1 and no 2. These are also the dominant groups in the matrix, in terms of size. The classical cluster no. 2 and no. 3 remained mostly unchanged, indicating a strong historical–thematic cohesion. Much less robust is the classical cluster 4, similar to no. 1, as its content has also been re-allocated between, primarily, the first and the last age-sensitive cluster (no. 1 and no. 6), but with less constituent elements than the first cluster, altogether.

So far, our quantitative comparison aimed at the relationship between what we called classical BC, and its age-sensitive modification (*asBC*). As emphasized above, our baseline for the classical BC was its unnormalized version, so that we could control for the sheer effect of the weighting procedure. Since, however, the most widespread application of bibliographic coupling is based on a normalized similarity measure, namely the cosine similarity of the respective reference sets, it is natural to ask how our proposed modification relates to the cosine-normalized version. In addition, even if we restricted our exercise to the set of unnormalized measures, a reasonable follow-up question would concern the selection of the weighting scheme itself: since the presented weighting function is rather specific, one would wonder the consequences of choosing a different one, such as a simple linear weighting, as a rather straightforward option.

Although this pilot study is intended to focus mainly on the potential role of *asBC* in revealing the "phylogeny" of a scientific problem (and not to test it against alternative methods in terms of precision or other general criteria for evaluating clustering schemes), we made a quick quantitative comparison of *asBC* with (1) classical, cosine-normalized bibliographic coupling based on cosine similarity (*cosine-cBC*), as well as with (2) a version of *asBC*, whereby the weighting function is a simple linear one (*linear-asBC*), respectively. This new, "semi-linear" weighting function (presented in Figure 5) assigns 0 to references previous to the XX. century, and a (linearly) age-proportional value between [1, 100] to the rest. Given the extended set of similarity matrices obtained with these modifications, we ran the same clustering procedure outlined above to arrive at two additional clusterings of our document set corresponding to *cosine-BC* and *linear-asBC*, respectively.

As a quantitative measure to contrast the results of the four clusterings (that is, *cBC*, *asBC*, *cosine-cBC* and *linear-asBC*), we used the age distribution of references in the resulting clusters, motivated by the idea that the proposed

measure (*asBC*) is supposed to best differentiate between clusters of source documents based on the recency of their shared references ("common ancestors"). More specifically, for each clusterings we obtained the clusterwise *age distributions of characteristic references*, and also conducted an ANOVA for each clusterings to see whether their clusters significantly differ from each other in terms of reference age. By "characteristic reference" we meant references whose occurrence in the cluster exceeds a predefined threshold. For the weighted versions *(asBC, linear-asBC)*, this was a weight-threshold (CDM>100, see the next section on the quantitative characterization of clusters), and for the unweighted versions *(cBC, cosine-cBC)* a mild rule of having an occurrence frequency > 3. To complement this assessment, we also obtained the dendograms resulting for each BC-version. The dendograms and the clusterwise age distribution of references are presented in Figure 3 and in Figure 4 below, respectively.

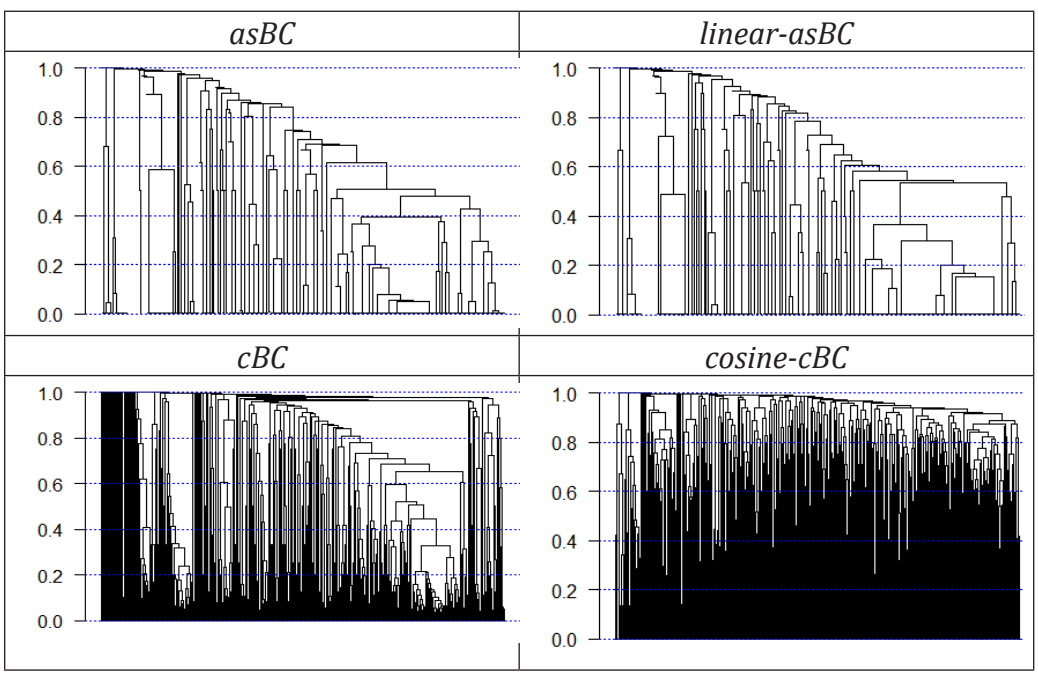


Figure 3. Dendograms obtained from the four clusterings

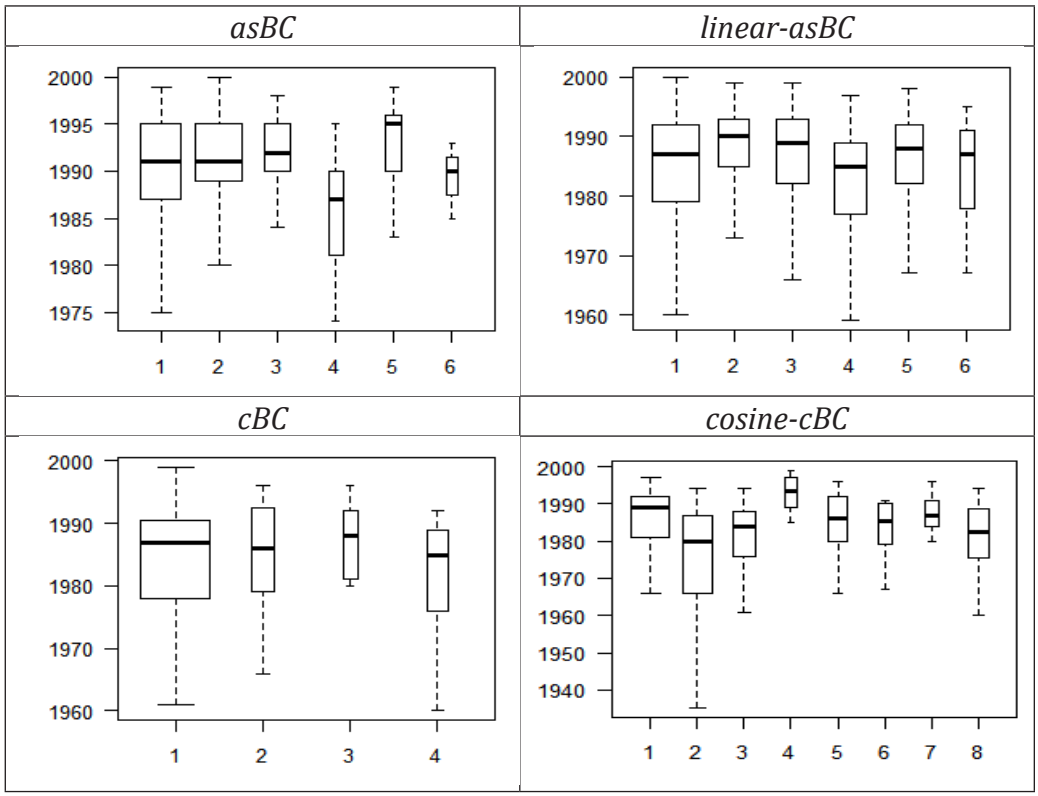


Figure 4. Clusterwise age distribution of characteristic references in the four clusterings

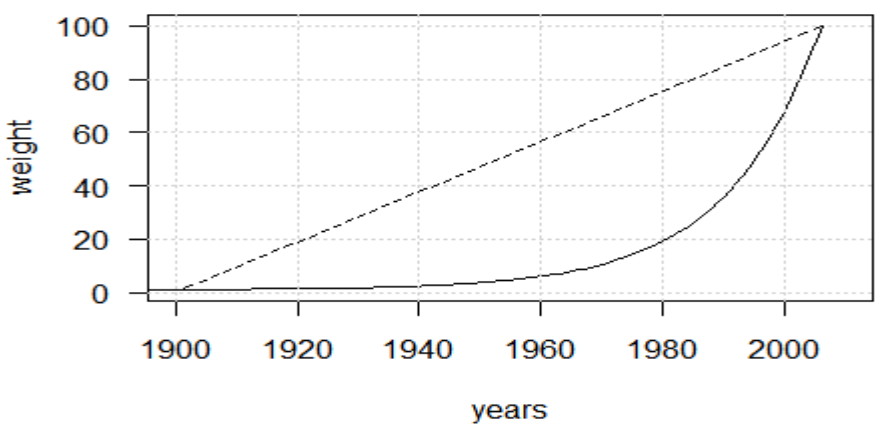


Figure 5. Comparison of the weighting functions for asBC (solid line) *and* linear-asBC (dotted line)

As one can tell via the boxplots, *asBC* shows a special distributional pattern as compared to any of the remaining versions. Cluster sizes strikingly differ (the area of the box is proportional to cluster size for each plots), and small clusters tend to occupy different periods (in terms of their distribution); the "last" cluster (#6) also has a rather small range. In fact, this is exactly what we expect from this method: to detect groups that are most tightly related by recency of their background, than others. The ANOVA, conducted selectively for the small clusters (to filter out the effect of "loose", less consistent groups) shows weak significance of these differences (The ANOVA results, with an indication of the groups included, are presented by Table 2. for all clusterings). Interestingly, the linear version (*linear-asBC*) exhibited high values for significance, but much less separation of the groups, or effect, is detectable from the boxplot. The other two versions (*cosine-cBC, cBC*) showed no significance (together with the selective group inclusion), and, more interestingly, these also did not seem to yield distinctive age-patterns for the clusters. The cosine version, however resulted in a more fine-graded structure (eight clusters).

Clustering	F	P(>F)	Sig	Classes
cBC	1.7549	0.1892	–	Class>1
asBC	5.1202	0.0273	*	Class>3
linear-asBC	12.521	0.0004	***	Class>1
cosine-cBC	1.4746	0.2269	–	Class>3

Table 2. Clusterwise ANOVA results for testing the differences of the age distribution of clusters

Dendograms reflecting the age-sensitive method can also be said distinctive when contrasted with the other two procedures. Both *asBC* and *linear-asBC* resulted in a more stratified structure: in both cases, small groups with higher proximities are iteratively nested into bigger groups having less tight internal proximities, while in the other two cases (especially in the cosine-based clustering), groups are better separated but with much less internal cohesion. Again, this difference is supposed to mirror the fact that the age-based methods are designed to best capture "phylogenetic" relationships with more and more similar nested groups, rather than some sort of general similarity.

In sum, based on this short comparison, age-base methods appear to suit the needs of a historical reconstruction of a topic. These are not necessarily better or more precise than other methods, such as the cosine-normalized BC, but the purpose of their application can be expected to differ. While the latter is best used for a high precision thematic structuring of a corpus, the former can better perform when a "phylogeny" needs to be reflected in the classification of a set of documents. A further advantage of the age-based weighting could be, via the possibility to parameterize the weighting scheme, to directly control the contribution of the age parameter to the result. However, the comparative study of *asBC* within the set of alternative

bibliometric clustering methods needs considerable future work, which is out of the scope of the present chapter.

Model 1.2: Multidimensional science maps

Based on the pool of various mapping methods described in Chapter 3, we are also proposing an integration of various maps to obtain a novel kind of science map we call *multidimensional*. The basic idea behind this proposal is to combine the most informative relations available from multiple maps based on different bibliometric indicators, in order to produce a rich structrue for the study of knowledge dynamics, with special emphasis on causal–historical connections. In particular, given any publication record *P*, our model consists of the set-theoretic union of three graphs extracted from *P*:

1) *Author-citation network induced by P.* The directed and weighted graph representing citation relations in *P* among authors within *P*.
2) *Keyword-citation network induced by P.* This rather unusual type covers the citation relations among key concepts within the corpus, based on the citation network of documents in *P*. In other words, this type of map is to reveal the descendancy of concepts and the development of the conceptual system based on actual knowledge flow. The network is a directed and weighted graph.
3) *Author-keyword network induced by P.* This map type differs, in terms of network theory, from both types 1–2 in that it is a so-called bipartite graph: it relates two different indicator set, that of authors and keywords. Practically, this bipartite graph creates a mapping betwenn the previous two network types, as it is to be induced by the author/keyword set within *P.*

To put it differently, our proposed model links or “matches” the knowledge flow among authors and concepts in a single representation via connecting the respective two graphs by a third one, that is a coupling of authors and

concepts. We argue that integrating these three bibliometric aspects of scientific discourses, or three traditional types of science maps has various benefits in the study of knowledge dynamics:

- *Semantically informative structure.* Traditional citation networks are, in most cases, difficult to interpret even if a tractable structure is detected in the graph. The primary reason is that widely used author citation networks speak of "formal historiography" in terms of (proper) names, therefore, interpreting the history requires additional sources of information on related concepts, ideas etc. In the network studied below, parallel citation networks induced by authors and concepts are linked together, ensuring a semantics for the analyst to author descendencies and interrelations identified in the graph, as a key to interpret underlying traditions.
- *Filling the gaps of missing links/data.* An inherent feature of "unidimensional" networks, especially in the case of keyword nets, that the underlying dataset is a partial one: in historical publication records, for example, older publications usually miss associated keywords or other content descriptors, typically due to a database/ indexer effect. When, however, connecting author and keyword nets, the complementer relation, that is, citations between authors, may fill the gap of missing citation links between concepts. Consequently, (historically) related sets of authors and terms may reveal themselves as cohesive groups to, e.g. community detection methods (see below), even in the absence of explicit relations on either side.
- *Historical (causal) relations instead of co-occurence.* A feature of high importance associated with the multidimensional maps is that is is constructed out of citation relations, that is, causal links in each dimension. Traditional concept maps are induced upon the co-occurence of keywords in documents, which is a useful indicator of

topics, but still an associative approach missing actual causal links or descendancy relations conveying the paths of knowledge flow. In our map, keywords a related through citation relations, allowing the analyst to directly track the evolution of the underlying conceptual system.

Our methodology in implementing the multimap proposal consisted of the following steps:

1) In the first step, based on a large-scale corpus collected in relation to the topic (see below), we obtained the three constituent map of the publication record, that is, the author–citation network, the keyword–citation network, and the author-keyword graph.
2) In the next step, the three graphs have been unified along common nodes (set-theoretically), resulting in the final, multirelational network.
3) We have filtered and normalized the raw multinetwork in a variety of ways, to adjust for the differencies between the traditional graphs. Most importantly, edge weights have been normalized to range from 0 to 1 in each constitutent graph, individually, since e.g. the frequent relations characteristic between keywords would have suppressed the much weaker associations in the author–keyword graph.
4) As the definitive step, in order to reveal the structure of the discourse, we have identified research traditions as subdiscoruses in the network as cohesive subgraphs via a community detection algorithm based on modularity maximization.

The method of community detection applied here is, in principle, the result of integrating two approaches aiding at community detection in complex

networks. The algorithm attempts to identify communities mostly based on the topology of the underlying graph, so that the resulting groups can be characterized as maximizing within-community connections, while minimizing inter-community connections:

(1) the Walktrap Community Findig (WCF) algorithm attempts to find dense subgraphs within a network by random walks (Pons & Latapy, 2005). The underlying idea for this algorithm is that short random walks with the probabilities determined by the edge weights are likely to circumscribe a community in the sense of being a set of densely and strongly connected nodes. The WCF algorithm works in an agglomerative fashion, starting with the strongest communities and merging the closest ones in consecutive steps until the whole network is reconstructed.

(2) The iterative procedure (1) is repeated until an optimal community structure is obtained. A now-standard method for optimization is the application of the network measure called modularity (Newman, 2006):

$$Q = \frac{1}{2m} \sum_{i,j} A_{ij} - \frac{k_i k_j}{2m} \delta(c_i, c_j),$$

where m is the number of edges, Ai,j is the corresponding element (weight) of the similarity matrix, k_i and k_j are the degrees of the corresponding nodes, c_i and c_j are the community indices the two node belongs to, respectively. $\delta(c_i,c_j)$ is a function that equals to 1 where both nodes are of the same community $(c_i = c_j)$, and 0 otherwise. Informally speaking, the function measures how "modular" a given network is under a certain partition of its nodes (community structure), i.e. how separated the different node types (communities) are from each other. Using this measure as the object function to be maximized, that is, by $Q \to \max$, the algorithm identifies the

optimal (most modular) partition of the network (without putting artificial constraints on CD, such as similarity thresholds).

Model 1.3: Knowledge diffusion through disciplines

Recent developments in the field of science mapping induced a variety of applications within evaluative as well as within structural scientometrics. The so-called *Science Overlay Map* technique or toolkit has been introduced by Rafols, Porter and Leydesdorff (2010), upon a mapping exercise of global science in terms of Web of Science Subject Categories or SCs (Leydesdorff and Rafols 2009). Via this toolkit, any collection of (WoS-indexed) publications can be represented as an overlay on the global map (hence its name), representing, therefore, its field composition and position on the scientific landscape. Consequently, countries, institutions, researchers, research topics or any other meaningful aggregations can be profiled and compared through this structural model.

An outstanding feature of the overlay toolkit is its capability to convey rich structural information on research profiles. The global science map (basemap) involved in profile mapping is a proximity network of research fields (SCs), based, in principal, on the bibliographic coupling of Subject Categories. Consequently, overlay maps for any aggregate of papers not only encode for the distribution of the aggregate over current fields of science, but also for the relation (cognitive distance) of the fields included in the overlay map. This nice feature is exploited with an ever increasing interest in applying this model within *Interdisciplinarity Research* (IDR). In IDR special focus is being put on inventing measures that summarize this multifaceted information within overlay maps in order to quantify the multi- and/or interdisciplinarity of research profiles. The most popular overlay-based measure of multidisciplinarity so far is the so-called generalized *Stirling index* (Stirling 2007) to be drawn upon any overlay map. Given a set of papers ranging over *n* Subject Categories, the measure takes the form

$$\text{Stirling index} = \sum_{i=1,j=1}^{n} p_i p_j d_j \text{ , whereby}$$

- p_i is the weight of the i-th Subject Category $i = 1, \ldots, n$,
- p_i is the weight of the j-th Subject Category $i = 1, \ldots, n$,
- d_{ij} is the distance of the i-th and the j-th Subject Category as determined by the basemap for the overlay.

The Stirling index can be interpreted as a measure of multidisciplinarity, capturing at least three aspects of cognitive diversity: the variety, the balance and the disparity of fields within research profiles (Leydesdorff and Rafols 2009). The index is rather flexible, as each parameter can be evaluated with different indicators: SCs can be weighted along by their relative frequency within the aggregate, but also with e.g. the impact of the associated papers; similarly, the distance term can be interpreted with a series of network measures, allowing for a variety of aspects to be quantified (Soós and Kampis 2011, 2012). The proposal discussed below highly depends on this IDR methodology: in what follows, by the *overlay toolkit* we mean both the *Science Overlay Map* model and the associated structural measures.

An overlay-based model of science dynamics

A huge potential in the approach outlined above is the application of the overlay toolkit in modelling the dynamics of science. In particular, two fundamental processes driving the development of scientific knowledge, namely, (knowledge) *diffusion* and (knowledge) *integration* seem to be outstandingly well characterizable via the overlay methodology. Diffusion and integration, in this context, are being conceptualized as converse processes: diffusion occurs when a subject propagates through a variety of research fields, yielding multi- or interdisciplinary composition of the research topic; integration, at the other extreme, is conceived as research undertaken in various fields converges towards a synthesis, often exhibited by the emergence of a novel field (for a detailed and conceptual discussion of these processes see Carley and Porter 2012).

Since the composition of a scholarly topic, in terms of research fields, can be x-rayed by mapping the body of related literature by the overlay toolkit, the evolution of its field composition can also be tracked via the same model. Therefore, diffusion and integration processes underlying the dynamics of the topic may be revealed and quantified as well, by applying – and adjusting, cf. below – the measures associated with the toolkit. The basic idea is to construct a "dynamic" overlay map for the subject, that is, a series of maps picturing field composition in consecutive time periods, whereby structural changes in the history of the subject become detectable through time.

Despite the clarity of this modelling scenario, some important choices should be made, in terms of bibliometric indicators, concerning the body of literature selected for instantiating the subject under study. Depending on this choice, different aspects of both diffusion and integration processes might be captured. The three basic options are summarized in Figure 6:

- The most straightforward method for grasping the topic dynamics is to partition related source documents into time periods (typically into years of publication), and subject each subset into analysis, separately. In terms of overlay mapping, annual changes in the field composition can be tracked within the corpus (type B dynamics).

- However, diffusion (and integration) may naturally be interpreted as being exhibited through citation relations, as basic indicators of knowledge flow. This aspect is best approached by a comparative analysis of source documents and citing documents: the overlay map of each annual set is, then, compared with the overlay map for the collection of docs citing those sets. In this case, the effect ("impact") of source documents on the scientific landscape may be directly observed, cohort by cohort (type A dynamics).

- The third basic type of relation to explore knowledge flow patterns is to focus exclusively on the citing side. In this case, a similar time series of overlay maps might be constructed as in the first case, but instead of source documents, cohorts of citing documents are being mapped in a consecutive manner (that is, each citing cohort determined by the respective cohort of source documents being cited). This third mode of exploration addresses a yet further aspect of subject evolution, namely the dynamics of the reception of the topic within the scientific landscape (type C dynamics).

Based on these considerations, an innovative work has recently been set forth by Carley and Porter (2012), demonstrating the use of the overlay toolkit in mapping science dynamics. Aiming at the quantification of the degree of diffusion/integration in knowledge transfer processes, the authors introduced the index of *forward diversity* grounded in the overlay toolkit. More precisely, they used the Stirling index as a diversity measure in a novel way to characterize the structure of knowledge transfer from scholarly fields. Schematically speaking, to explore diffusion processes (1) a group of Subject Categories (SCs) was selected from the Web of Science databases as representing benchmark fields and (2) the overlay map of the record of citing documents was obtained for each benchmark SC. Upon field composition, they characterized the citing side of benchmark SCs with the Stirling index, reflecting the intellectual diversity of research relying on (referring to) the filed (SC) under study. Quantifying structural diversity of the citing side provided a simple yet powerful formalization of the extent of knowledge diffusion originating from a certain field. The kind of knowledge dynamics addressed was what we hereby call “type A dynamics”: the diffusion process was followed along citation relations of source document cohorts.

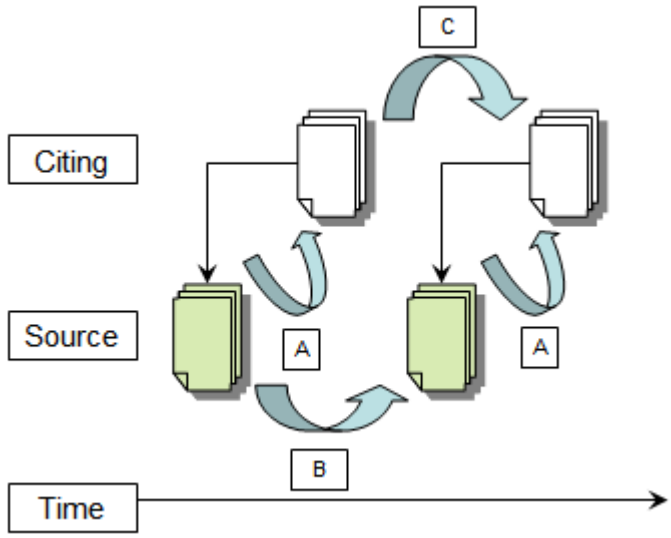


Figure 6. Dimensions of science dynamics typified via bibliometric relations.

A proposal towards dynamic diversity measurement

The proposal formulated in the rest of this chapter is closely related to the concept of *forward diversity* as a knowledge diffusion indicator, but is centering around an aspect of knowledge dynamics that has seemingly been neglected so far whenever the overlay technique was adopted. Simply put, the extent of knowledge diffusion has always been approached in an "assymetric" way. Diversity values (the Stirling index) has usually been calculated for the resulting field composition, independently of the initial one. Given e.g. "type B dynamics", obtaining diversity values for consecutive document cohorts yields a time series of diversity values (one for each timeslice), which shows how diversity changes step by step, but tells little about the underlying change of field composition. Note that the same degree of diversity may be produced by a totally different set of SCs, so that it is logically possible (though, of course, empirically rare or even unrealistic) that no (diversity) change is detected in a moment of "revolutionary transition" within the diffusion process. Similar considerations apply to the comparison

of the cited vs. the citing side within a body of literature (type A dynamics): diversity on the citing side alone may inform us on the scope of knowledge reception, but to grasp *diversification* (vs. diversity) in the diffusion process, a comparison with source documents (the cited side) should also be somehow involved into the method of measurement.

In general, in order to detect genuine knowledge diffusion (either between time periods or within knowledge flow relations) the change in field composition between the source and target set seems to need some treatment. In terms of overlay mapping, the *distance* between source and target maps should be accounted for to track the shifts in the SC configuration of the topic. As the authors themselves point out, the index of *forward diversity* indicates zero diffusion when both the source and the target document sets are assigned to a single Subject Category, respectively – even when the two SCs are completely different ones. Being a definite case of field shift, one might consider it some undesired feature of the measurement, which is exactly the starting point of the present study.

To overcome this difficulty, it is tempting to turn to a tradition of measurement that is being entertained in the strongly related area of technology mapping. *Technological distance* is a concept central to the study of technology dynamics, when the basic problem is conceptually identical to the present one: to formalize the extent to which two technological profiles (of markets, of firms, or of patent portfolios etc.) differ structurally (cf. Los 2000). In technical terms, technological distance is mostly formulated as the mathematical distance of two vectors, – based on some distance metric – each representing a profile, that is, a distribution of products, patents, services over technological categories (e.g. patent classes). The main advantage of this model, as compared to diversity measurement, is that vector distances reflect an “item-by-item comparison” of profiles, therefore, the actual change in the composition of portfolios is better accounted for.

On the other hand, simple portfolio vectors lack the rich variety of structural information that overlay maps (and the related diversity measurement) possess, as they do not mirror the proximity of categories themselves in the first place. Therefore, whether the difference between two distributions (i.e. technological profiles) involves redistribution among distant or close technologies cannot be told from this model alone.

Upon these considerations, our main aim here is to make an attempt to combine the overlay methodology, utilizing its rich measurement potential, with the concept of technological distance for measuring knowledge diffusion. Our strategy is, then, to construct an overlay-based measure of knowledge dynamics that incorporates (1) the change in the field composition between the source and the target set, respectively, and (2) the extent of potential diversification this change produces within the field composition. Our proposal to meet this challenge is what we refer to as "dynamic diversity measurement".

Mean Overlay Distance (MOD)

The basic idea of a dynamic diversity measure is to formulate a distance measure between two overlay maps representing the source document set and the target set, respectively. This measure is expected to take into account the similarity *between maps* and the distances *within constituent fields of the profiles* simultaneously, the latter in terms of the science map. That is, firstly, we are interested in how much the overlay map is being restructured between the source and target, and, secondly, at what distance the new field composition lies from that of the source, according to field distances indicated by the basemap.

As the previous studies, we also utilize the Stirling index to capture this two aspects of knowledge diffusion. The crucial difference is that, instead of

characterizing single overlay maps, we apply the measure to the *comparison* of two maps, one of which is of the source, and the other is of the target. Given a source document set characterized with N = *n* research fields (Subject Categories), and a target set distributed over N = *m* Subject Categories, the proposed measure, ***M**ean **O**verlay **D**istance*, can be defined as follows:

$$\text{MOD} = \frac{1}{n^{*}m} \sum_{i=1,j=1}^{n,m} p_i p_j d_j \text{ , whereby}$$

- p_i is the relative frequency of the *i*-th Subject Category within the **source** SC-profile, *i* = 1, … , *n*,
- p_j is the relative frequency of the *j*-th Subject Category within the **target** SC-profile, *j* = 1, … , *m*,
- d_{ij} is the distance of the *i*-th (source) and the *j*-th (target) Subject Category as determined by the (common) basemap for the (both) overlays.

As can be seen from the definition, the MOD index operates on two maps the same way just as the Stirling index operates on a single map: it imposes a pairwise comparison of source and target fields (SCs), and favors those pairs, that are significant within the respective map (has high share among constituent fields), and, at the same time, cognitively distant from each other. The calculation can be conceived as the summation over the cells of a matrix of weighted source-SC-by-target-SC distances (Table 4). In other words, MOD measures both the overall (structural) difference and the (cognitive) distance between two maps. Therefore, while the previous use of the index on single overlays reports the *diversity* of SC composition, this "dynamic" extension adds the *diversification* occurred between two maps. In order to control the effect of size, the value of the "dynamic" Stirling index is normalized by the first term of the MOD formula, yielding an average of weighted distances between the two maps (hence the name *Mean* Overlay Distance).

Target Source	**$SC_{target-1}$**	**(…)**	**(…)**	**SC_m**
$SC_{source-1}$	$P(SC_{source-1}) \times p(SC_{target-1}) \times d(SC_{source-1}, SC_{target-1})$	(…)	(…)	(…)
(…)	(…)	(…)	(…)	(…)
SC_n	(…)	(…)	(…)	$p(SC_m) \times p(SC_n) \times d(SC_m, SC_n)$

Table 4. Source SC x Target SC matrix underlying the MOD index

Overlay Diversity Ratio (ODR)

Given the strategy of a comparative use of the Stirling index, it is of outstanding interest how the application introduced above performs against previous uses in empirical settings. More precisely, the question is whether any process of knowledge diffusion – being modelled via the overlay methodology – shows a different picture when operationalized via the diversification-oriented MOD index versus single-map based diversity. This previous use of the Stirling measure, to simply distinguish terminologically from the MOD index, we may call ***O**verlay **D**iversity* (OD).

A conceptual difficulty in such a comparison lies in the very fact that the MOD index is designed for between-map usage, while the OD index applies to within-map assessment. A direct contrasting of the two measures requires a further step, whereby the *usage* of the two indices both describe the same phenomenon, namely, the transition of field composition within the evolution of a subject matter. Also, we intend to keep the original properties of the OD index for a meaningful comparison.

To meet these requirements, we introduce the concept of *Overlay Diversity Ratio (ODR)*, which is nothing more than the ratio of diversity values (ODs) for the source map and the target map, respectively. That is

$$\text{ODR} = \frac{\boldsymbol{D}_{target}}{\boldsymbol{D}_{source}}, \text{ whereby}$$

- OD_{target} is the Overlay Diversity of the target set (as measured by the Stirling index),
- OD_{source} is the Overlay Diversity of the source set (as measured by the Stirling index).

In verbal terms, the ODR index accounts for the relative change of the diversity in field composition between two maps. Its value equals ODR=1 in case when the transition does not affect the degree of diversity. If ODR > 1, the transition leads to an increase of diversity (a potential indication of knowledge diffusion), ODR < 1 reports a lower degree of diversity after the transition (a potential indication of knowledge integration).

Application to the corpus

With the extended overlay toolkit discussed above, we have applied the dynamic Stirling measure for quantifying knowledge diffusion throughout the Species Problem, and, consequently, to gain insight into the interaction of consituent fields and disciplines. Just as the proponents of *forward diversity*, we also addressed citation relations to track the potential diffusion process. In particular, (1) annual cohorts of the selected publication record were obtained, and (2) for each cohort, all papers citing its members were collected. In terms of our typology, the case study concerned **type A dynamics** (knowledge flow between the cited and the citing side). However, in order to capture the overall evolution of the topic, beyond annual sections, – that is, pairs of source cohorts and citing paper sets – we also profiled knowledge flow in a cumulative manner, by aggregating source cohorts up to each year along with the papers citing that aggregate. The rationale behind this perspective is to allow the MOD index to capture the annual extent

of knowledge diffusion relative to the prehistory of the discourse at each time period, not only to the extent characteristic of a particular time period (based on the publications originating therefrom). As a consequence, with this cumulative method, both the annual values of the *Mean Overlay Distance* and the *Overlay Diversity Ratio* implicitly incorporated the measurement of **type B** and **type C dynamics** as well, inasmuch the development of field composition were captured via time-aggregation at both the cited and the citing side.

To set out formally, we have combined the above measures and methods in the following arrangements:

- *Diversification from annual sections.* For each year within the time coverage of our sample the overlay maps for annual cohorts and the related citing papers were constructed. On this basis, the *Mean Overlay Distance* between the two maps per year was obtained for monitoring the dynamics of knowledge flow in a cross-sectional perspective (that is, in each time period separately).

- *Diversification by each year (cumulative approach).* Pairs of cited–citing overlay maps were also generated by the cumulative method. In this case, the cited side for any year *Y* was translated into an overlay map of the group of sample documents published in the year $y \leq Y$. As a consequence, each overlay map contained that of the previous years at both the cited and the citing side. It follows that annual maps (as compared to predecessors) showed the new developments (new fields) for each year in the history of the topic. The MOD index was also calculated upon this series of map pairs.

- *Diversity change by each year (cumulative approach).* In order to contrast diversification with diversity change, the *Overlay Diversity Ratio* was also applied for the cumulative (or "historical") series. In

particular, the ratio of the Stirling index for the cited and the citing side maps was obtained in each time period, based on the time-aggregated maps.

Chapter 6: A methodology for latent conceptual organization

Complementing the citation-based methods discussed in previous chapters, we have also developed a rich toolkit to explore the latent conceptual development of the Species Problem along the timeline.

As we have seen above in the case of dynamic overlay maps, this science overlay map approach triggered a complete analytical framework (*ScOM;* Rafols et al 2010) . The core idea is that any kind of S&T actor (individual, group, institution, country, region etc.) can be modelled by overlaying its research profile on this global science map. Beyond "x-raying" research activity, various structural properties of research profiles can be studied and even quantified based on the underlying science map, such as patterns of inter- and multidisciplinarity – this is why a toolbox of ScOM-based measurements has been developed within interdisciplinarity research or "IDR" (Rafols and Meyer 2010).

The main aim of the chapter presented below is to elaborate a multi-purpose framework, inspired by Science Overlay Mapping, aiming at the analysis of the latent cognitive (thematic or topical) organization of scholarly discourses. The backbone of our proposed framework is identical to that of ScOM: (1) Identify the main topics in a scholarly corpus, (2) Draw a global "discourse" map showing its internal cognitive organization by the interrelations of topics and (3) build an analytical toolbox to overlay any parts of the corpus on the global map, and quantify its cognitive patterns by exploiting the undrelying map. The resulting analytical framework (Topic Overlay Mapping, hereafter: TOM) naturally lends itself to three main applications: (1) the measurement of cognitive complexity of papers or paper sets (2) the comparison and clustering of papers based on their overlay maps, and (3) the measurement and visualization of the cognitive dynamics and development of the whole body of literature under study.

Science overlay maps (ScOMs), as introduced by Rafols and Leydesdorff (Rafols et al. 2010), are based on a complex network of research areas (WoS journal categories, in this case), drawn from their respective proximities in terms of referencing behavior (citation patterns) – called the *basemap* (of science). Such a global map of science is then used to represent research profiles from any set of WoS-indexed papers, showing how those papers are distributed over the basemap – resulting in the *overlay map*. Upon this model, a rich analytical toolbox can be developed to quantify structural properties of a research profile (cf. Rafols and Meyer 2010, Soós and Kampis 2011, 2012). In close analogy with this approach (or as an extension of the method), our framework (TOMs) consists in the following modules:

The construction of Topic Overlay Maps

The central concept in our model is the Topic Overlay Map (TOM), designed to represent the position of any aggregate of papers within a rich cognitive map of a scholarly discourse. Given a corpus (or a set of bibliographic metadata) associated with a scientific discourse, a TOM is created in four steps:

(1) A term-proximity (weighted) graph is obtained from the whole corpus (based on the joint distribution of textual descriptors, e.g. keywords, within the corpus).

(2) The graph is clustered (with an appropriate community detection algorithm) into cohesive term sets, as proxies for main topics of the discourse.

(3) As the key step, upon the preestablished cluster structure, a new graph is constructed that captures the relationship of clusters emerged from the simple term network. Formally, this graph is a proximity network of clusters (as nodes), based on the connectedness of their respective elements (terms) in the underlying term graph. As this new structure is supposed to formalize the cognitive organization of the discourse by a weighted proximity network of topics (proximities expressing topical interrelations), it serves as the basemap in our model.

(4) The final step consists of overlaying a selected set of documents (belonging to the corpus) on the basemap. This can be done by remodelling documents in terms of the clusters of the basemap. Each paper or paper set is characterizable via the distribution of its textual descriptors over topics (the cluster set). This distribution can be visualized on the basemap (cf. subsection C), or used quantitatively in map-based structural measures (cf. B), and is called (along with the map properties) the overlay map for the paper set.

Quantifying structures of cognitive organization

A peculiar feature of ScOMs for the study of research profiles is that, contrary to simple distributions which allow for measurements of the variety and balance of constituent research fields, science maps also provide indicators for the disparity of the profile, i.e. the cognitive distances between research areas involved in the profile. It is a benefit inherited by Topic Overlay Maps as well, which can be exploited by applying structural map-based measures originally proposed to quantify the diversity (multidisciplinarity) and cohesion (interdisciplinarity) of research profiles. We developed our framework to involve the following applications:

(1) Quantifying cognitive (topical) complexity of publication(s). The so-called Stirling index (Stirling 2007) is applicable to a topic overlay map (cf. Rafols and Meyer 2010), measuring both the topical composition and its cognitive scope for publication(s), indicating the underlying topical complexity. The cognitive scope is based on the distances of constituent topics within the basemap.

(2) Similarity of overlays and document clustering. The generalization of the classical Cosine similarity measure, called the Proximity Weighted Cosine Similarity (Zhou et al. 2012) is applicable to compare any two topic overlays. Beyond the similarity of papers in topic composition, it is also sensitive to the cognitive proximity of any two papers (overlays). This similarity

measurement lends itself to assist a document clustering that results in paper groups reflecting complex relations or "functional overlaps" between topics rather than their simple combinations.

Visualizing the cognitive structure and dynamics

A third application of the framework of TOMs lies in its capacity for visualizing the cognitive organization of discourses, along with their dynamics. Topic Overlays are, in the primary sense, visualizations of a particular topical portfolio (of a paper set) upon the basemap, that reveals the relatedness of topics as well. The visualization consists of customizing the basemap via setting node sizes to express the contribution of each topic (node) to the portfolio, node size being proportionate to the degree of topic contribution. In such a way, the evolution of a discourse can also be set out visually by overlaying the underlying corpus partitioned into consecutive time periods: the dynamics of topics and their interactions as well as their development can be tracked throughout the history of the discourse.

Term graph of textual descriptors: the map of the species problem

Implementing step 1–2 of building a basemap for the discourse, a term graph was devised from the core bibliographic record. As textual descriptors of papers, author keywords, title words and also keywords extracted from the reference list of papers were selected. To obtain the terms, this set of words was normalized with NLP procedures (stemming etc.) A standard similarity network of the most frequent terms was obtained upon the term–document matrix (via the Cosine similarity). Finally, the resulted graph was subjected to a community detection algorithm sensitive to edge weights (modularity maximization via random walks). The procedure yielded in 14 topics. The term graph along with the detected topics (clusters) is depicted on Figure 7.

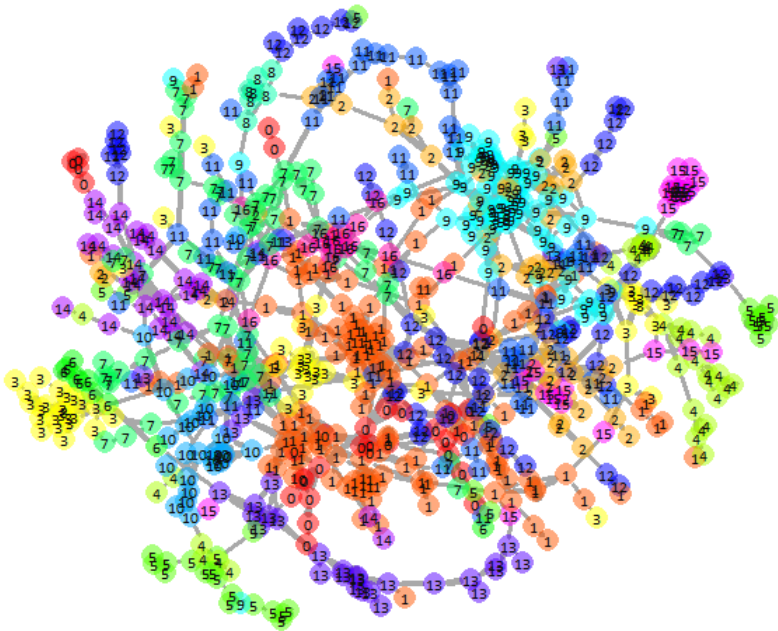


Figure 7. The term graph and topics (numbered and colored) of the discourse. Only the giant component is shown.

Basemap of topical relations

To obtain the basemap (step 3), the relatedness of topics was estimated based on their connection patterns in the term graph. In particular, the "overlap" between two topics A and B was defined as a weighted average of the weight of edges connecting the elements of A and B. From running this measure on the term network, a proximity graph of the 14 topics was drawn, yielding the basemap of the discourse for the whole period under study (Figure 8)

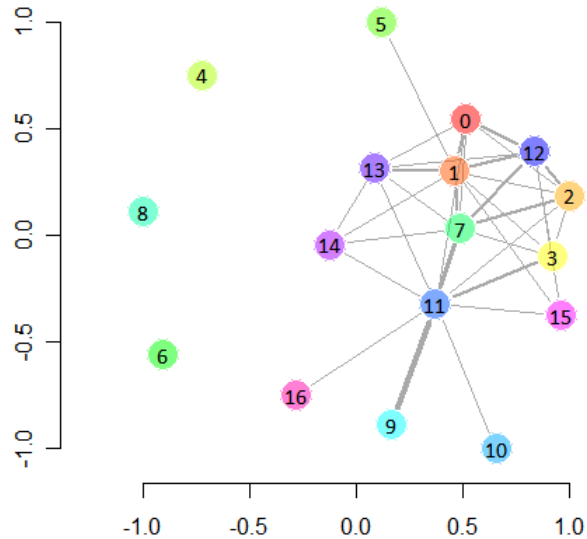


Figure 8. Topic basemap constructed from the clustered term graph (with the same cluster numbering and coloring). Only the links with strength above a pre-selected threshold are shown.

Trends in the history of the species debate

Given the basemap of the selected theme, we used the TOM-based document clustering method to reveal the overall conceptual organization of the Species Problem. Based on the topic overlay map of each paper, after (Zhou et al 2012), we clustered them via their pairwise proximity weighted cosine similarity (cf. section II.B, part 2):

$$\varphi(X,Y) = \frac{\varphi_{XY}}{\sqrt{\varphi_X \ \varphi_Y}},$$

$$\varphi_{AB} = \sum_{i,j} S_{A(i)B(j)} p_{A(i)} p_{B(j)}$$

whereby , and

S_{ij} is the proximity of categories A(i) and B(j) within the basemap.

$p_{A(j)}$ and $p_{A(j)}$ are the relative share of topic i and j in paper A and B, respectively.

The clustering was achieved by subjecting the resulted similarity matrix to a hierarhic (agglomerative) clustering algorithm (average linkeage clustering).

Comparing the TOM-method against the VSM-based clustering

As a preliminary assessment of the performance of TOM in our application, we evaluated our TOM-based clustering of the corpus against a classical document clustering method based on the Vector Space Model (VSM). To that end, we also subjected our corpus to a second clustering, whereby paper similarity was established by the standard cosine similarity measure, based on a term–document matrix normalized by the tf–idf method. The grouping of documents was obtained by the same agglomerative clustering algoritm, as in the TOM-based case.

A rather straightforward comparison between the two approaches is to contrast the resulted cluster structures. The dendogram from the TOM-based clustering, and the VSM-based clustering is presented in Figure 9 and Figure 10, respectively. The difference of the two dendograms is rather striking: while the TOM-based dendogram conveys a strong internal structuring of the corpus, the VSM-based dendogram mirrors a lack of any proper structure (in fact, this cluster tree was cut at a very high level – that is, extremely low level of relatednes – to even make the graph readable). It suggests that while the standard method was not able to find coherent-enough topics within the discourse, the TOM-based approach could discriminate between topics and even subtopics at a seemingly efficient way.

Since we were primarily interested in the science mapping potential of the TOM toolbox, instead of further quantitative measures, we have focused on the qualitative comparison of the two clusterings, concerning the differences of how the corpus is being organized (what kind of trends are being revealed) according to each, respectively. To that end, we have obtained the set of clusters from both dendograms most naturally mirroring their internal structure. Instead of cutting the trees at a pre-defined height, we used the

so-called “dynamic tree cut algoritm”, that detects the clusters depending on tree shape (http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/BranchCutting/).

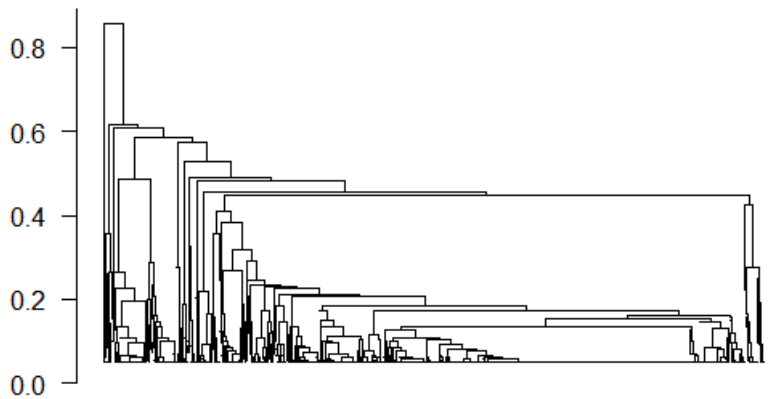


Figure 9. The cluster structure resulting from the TOM-based clustering of the corpus

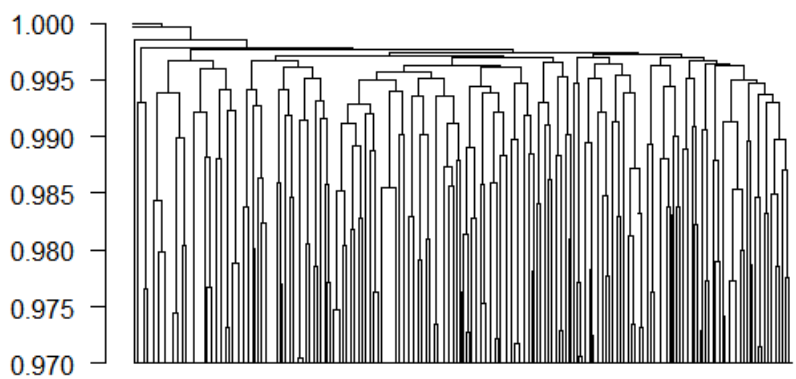


Figure 10. The cluster structure resulting from the VSM-based clustering of the corpus

The procedure resulted in approx. 20 topical clusters in the VSM-case, and around 50 topical clusters in the TOM-case. To establish a qualitative relation between the two partitions, a keyword-profile was generated for each group

in both case, showing the distribution of their most frequent characteristic concepts (author keywords).

The main finding from contrasting TOM-topics and VSM-topics by the overlaps of their keyword profile was consistent with our preliminary expectations. In general, (1) TOM-topics were mostly orgaized around general or "leading" concepts or research subjects indicating many particular lines of investigation that are connected to the general subject, while (2) VSM-topics were more narrow in scope, centered around those particular lines or subtopics. In other words, while VSM.-topics mirrored the "table of content" of the discourse, TOM-topics combined the related "chapters" into single clusters, clearly highlighting the – usually latent – research subject that was central to those chapters.

clusters	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
1	14	3	2	9	5	8	2	5	4	2	2	5	0	1	2	1	1	2	2	3	2	1	1	2	0	1	1	2	2	0	0	1	1	1	1	0	1	1	0	2	0	1	0	1	0	1	3
2	17	8	2	1	6	2	0	2	6	0	1	2	0	1	3	0	1	3	1	2	3	3	0	2	1	1	0	4	2	0	7	2	2	5	5	0	0	3	0	2	0	0	0	1	0	0	1
3	10	0	0	2	0	0	1	0	2	15	0	2	0	0	3	1	0	4	0	3	3	1	0	2	1	5	0	1	3	0	7	1	2	0	0	2	1	0	17	4	0	0	0	2	2	0	0
4	24	2	0	0	9	1	0	1	2	0	0	2	0	5	2	0	1	2	0	0	3	3	0	4	0	1	3	2	0	0	4	5	1	3	4	1	0	2	0	1	0	0	0	3	0	5	0
5	12	1	1	0	0	1	28	0	1	0	0	0	33	1	2	1	0	0	1	1	0	0	0	1	2	0	5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	1
6	27	6	0	0	3	0	1	0	0	0	0	1	0	4	3	0	21	0	1	0	0	3	3	0	15	1	0	0	1	0	0	3	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0
7	26	2	0	0	2	2	3	5	3	0	0	3	2	0	0	2	3	0	17	3	0	6	0	0	5	3	3	2	2	0	0	2	0	2	2	0	0	3	0	0	0	0	2	2	0	0	0
8	11	3	23	0	2	0	0	2	0	0	3	5	2	0	2	10	0	2	6	0	0	2	0	0	2	2	2	3	2	0	0	2	2	0	2	0	0	2	0	3	6	0	0	0	0	0	3
9	20	2	2	0	2	0	3	0	0	17	0	2	0	0	0	0	0	0	0	2	0	0	0	3	3	3	2	2	0	0	2	0	7	0	0	5	3	0	7	3	0	0	2	3	2	5	0
10	9	7	4	2	0	0	2	0	7	0	2	0	2	2	2	0	0	2	0	0	0	2	16	0	0	0	14	0	2	2	2	4	0	0	4	0	0	5	0	0	0	0	2	5	0	5	0
11	12	4	0	2	0	0	2	0	2	0	21	2	0	19	12	2	4	0	0	0	0	0	0	2	0	0	0	0	0	2	4	0	2	0	0	4	0	2	4	0	0	0	0	0	0	0	0
12	31	15	4	0	0	0	0	6	0	0	2	0	6	0	0	2	0	8	0	2	0	0	8	0	0	0	2	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	2	0	0	0	4
13	16	4	4	0	0	0	0	2	0	0	0	6	0	0	4	4	0	12	2	4	0	0	0	2	0	8	0	0	0	2	4	2	4	0	0	0	0	6	2	0	0	2	0	2	2	2	4
14	11	6	15	0	2	4	2	4	4	0	2	0	0	2	2	13	0	0	2	0	0	2	4	0	0	2	0	0	0	0	0	0	0	2	2	0	0	0	0	2	13	0	0	0	2	0	0
15	13	2	19	0	0	0	0	0	0	0	6	0	0	0	0	23	0	0	2	4	0	0	0	2	0	2	0	4	0	0	0	2	0	2	2	2	2	0	0	2	9	0	0	0	0	0	0
16	29	2	0	0	2	0	2	4	2	0	11	0	0	7	0	0	0	2	0	0	0	11	2	2	0	0	0	0	0	0	0	0	4	0	0	0	11	0	0	0	0	0	0	2	0	4	0
17	29	5	2	5	0	0	2	5	0	0	2	0	0	0	0	0	0	5	0	2	0	2	0	5	0	0	0	0	0	12	0	0	2	0	0	0	2	0	0	0	5	2	5	0	7	0	0
18	3	0	0	0	3	0	0	0	3	12	0	3	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	12	0	0	0	9	0	0	0	3	0	0	0	0	0	29	0	0	0
19	18	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	0	3	0	0	3	3	0	0	0	0	0	0	24	0	0	0	0	0	0	6	0	0	0	0	30	0	0	6	0	0
20	12	0	3	0	0	3	3	3	3	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	12	0	0	0	0	0	0	0	0	0	0	0	3	15	9	0	27	0	0
21	16	0	0	0	0	3	0	0	0	6	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	0	16	3	0	0	0	0	0	0	0	0	42	3	3	0	0	0	0	0	0	0	0	0

Figure 11. Profiles of individual TOM-clusters (rows) in terms of the VSM-clusters (columns). Cells indicate % of row totals.

We can illustrate this tendency by exploring the content of a particular set of clusters that are clearly related, as being wittnessed by the cross-tabulation of the two clusterings shown in Figure 11, throughout the TOM- and VSM-based case. The TOM-based topic shown in Figure 12 is characterized by the theme of modelling and explaining the behavior of ecosystems (TOM-cluster 5). On Figure 13, the two VSM-clusters corresponding to this TOM-cluster

are depicted (VSM-clusters 6 and 12). The topics connected in the previous case under the theme of ecological modelling are clearly recognizable in the two clusters, but being sharply separated into different and, in this sense, unrelated research directions, one focusing on the study concerning the role of species in ecosystems (keystone species problem, foodwebs etc.) and the other addressing the research on the mathematical modelling of such roles (Lotka-Volterra models, equations, asymptotic stability etc.). That is, the TOM-methodology seemed to be able to recognize the linkeages between narrower topics, and combine document sets so that these linkeages became visible. In sum, the results show that, in comparison with standard document clusterings, the overlay-map-based method leads to a much more informed grouping of papers into research lines, mirroring the dominant interrelations of basic topics.

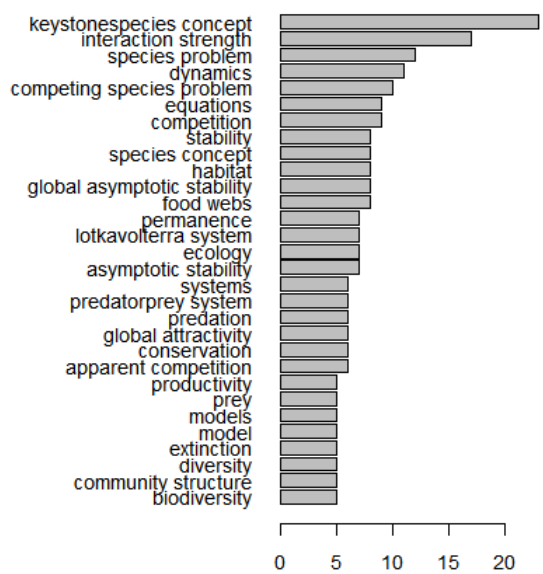


Figure 12. The "ecosystems"-related cluster from the TOM-clustering

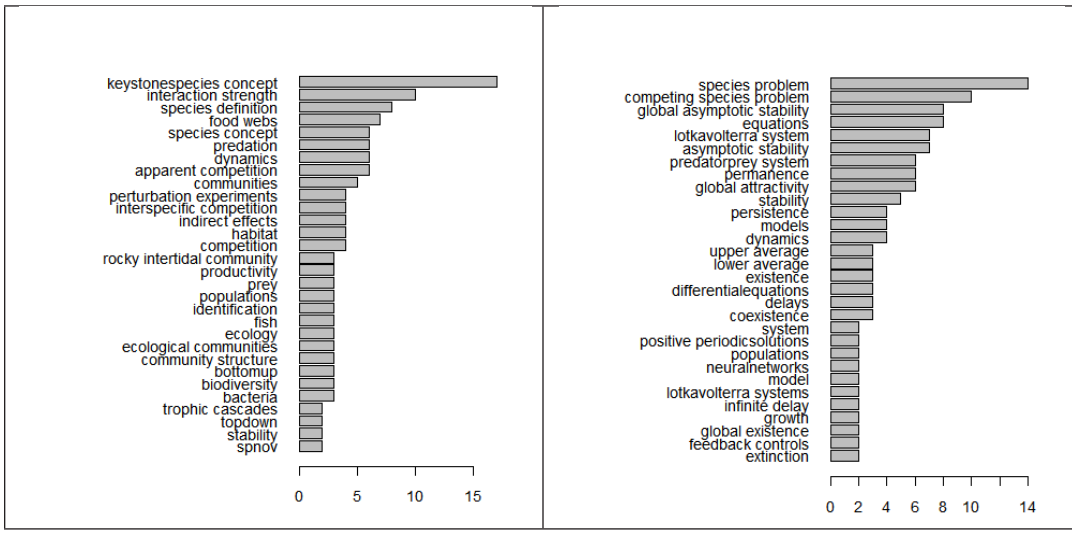


Figure 13. Clusters overlapping with to the "ecosystems"-related TOM-cluster based on the VSM-clustering

Chapter 7. Results and discussion

With the inventory of science mapping methods elaborted in previous chapters we now turn to the results of their application, that is, to what they reveal about the organization of the species debate (RQ1). This will also enable us to conduct the qualitative validation of our instruments (RQ2a) in a parallel fashion, that is, via a "constant comparison" of our results with the "standar story" provided in Chapter 1, during the discussions below. In the following sections, the reconstruction of the problem will proceed by evaluating the picture emerging from each application. In particular, the (1) historical subdiscourses identified by the *Age-Sensitive Bibliographic Coupling*, (2) the conceptual relations based on the citation network addressed by the *multimap approch* and (3) the latent conceptual structure and its development – the results from *topic overlay mapping* – will be exposed. As a fourth assessment, (4) the interaction of disciplines and fields will also be discussed via the *dynamic overlay map* approach.

Results based on model 1.1 : Uncovering historical subdiscourses

Before turning to the historically informed structure revealed by the age-sensitive version of bibliographic coupling, some additional methodological remarks should be provided.
According to our primary interest in applying the *asBC* method to the historical corpus in the focus of this study, we also investigated the content of the resulting document clusters, in relation to the classical ones (resulting from simple bibliographic coupling). To this effect, we followed a strategy based on two pillars:

1) Since mapping the intellectual structure of the topic was modelled via references, for the qualitative–narrative characterization of these clusters we also relied on the contribution of references to the formation of clusters.

2) In order to obtain a mapping in an economic way, that reveals both the profile of the new clustering and the difference between the "old" and the "new" profile, we did not aim to describe all groups. Instead, we selected a set of clusters that best represented these two aspects at once.

Point 1) above has been addressed by the following procedure: for each selected cluster *C* the references of documents belonging to *C* were collected and ranked, according to their cumulative weight in *C* (that is, their weight used by the *asBC* method times the number of documents they referred by, within *C*). Note that such a cumulative weight is proportional to the contribution of the particular reference to the formation of *C*. In other words, this ranking shows how important a particular reference in the intellectual background of *C* is. Based on this ranking, we obtained the first *n* most important reference in *C* to draw the profile of the cluster. The threshold *n* was based on a "knee plot" of ranks: the weight-based ordering of reference sets in each case led to a typical powerlaw-like curve with a relatively few references – with high cumulative weight – playing a major role, and many more contributing to a much lower level in itself. We identified these highly-weighted refs as residing in the first, most rapidly ascending section of the weight-curve that ends with a change of slope, the so-called "knee" that can be seen as a transition to the almost flat section of the curve. As the most important descriptors of *C*, we called this *n* references (above the knee of the curve) as the *core* of *C*. In what follows, beyond its description, the core is presented for each cluster under consideration as a set of references, and supported by the knee plot of the cluster. The knee plots are presented under Figure 15. Core references are also included in for each cluster, in the form of ranked lists, collected in Table 3.

Point 2) of our strategy was achieved by selecting *asBC* clusters no. 1–4 to look after contentwise, together with their two sublcusters. One of these

is (1) the fragment of no. 1 that previously belonged to the classical (*cBC*-) cluster 1, referred to as 1/1, and (2) another fragment of no. 1 that previously was part of the classical cluster 4, referred to as 1/4. The explanation of this choice leads back to Table 1. It can be seen that by examining *asBC* clusters no. 1 and 2, we can gain insight to the two dominant clusters (in terms of size) of the new thematic profile. On the other hand, since the vast majority of the first *cBC*-cluster has been reallocated between these two and, in addition, no. 4, we may also observe how the oversized "old" thematic group (no. 1) has been reconceptualized by the age-sensitive method. The two sublcusters 1/1 and 1/4 further refine this picture, as while new clusters no. 2 and no. 4 were born almost exclusively from the classical no. 1, new cluster 1 also inherited from old cluster 4. Finally, new cluster no. 3 is discussed as left rather intact (being almost identical to old cluster no. 2). In sum, by this selection, both novel and unchanged parts of the new profile are sampled (*asBC*-clusters 1–2–4 and 3, respectively), and also the relation of the two clusterings may become visible.

Based on these considerations, the historically informed structure of the species problem can be described with the following profiles:

- Cluster no. 1: the BSC and the debate over the theory framing the species concept

The core of the first cluster contains approx. 50 important references, ranked with their cumulative weights in Table 3[1]. (the knee plot on Figure 15 suggested a threshold of cumulative document weight, CDM > 150). Highly-ranked references are the position papers on the species concept since the modern synthesis. Most striking, especially from the full list of core references including books and book chapters as well, is the dominance of Ernst Mayr,

[1] In Table 2 only journal publications are demonstrated, therefore, the actual number of references included in the table is smaller than the size of the whole core.

the champion of the "biological species concept" or the BSC (cf. Mayr, #4) what, basically, launched this debate in the context of the synthesis. Several position papers, upon debating the BSC, ranked high in this list. These papers are also classical proposals of infamous alternative species conceptions (not just definitions), such as the "pluralistic conception" or "species pluralism" (Mishler, #10), the "evolutionary species concept" (Wiley, #13), the "genetic species concept" (Masters, #15). With somewhat lower weights, but two further definitions also exhibit themselves, namely, the "phylogenetic concept" (Nixon, #18), and the "ecological species concept" (van Valen, #30), though the latter having the lowest rank in the list.

Beside the collection of proposals to challenge the BSC as the concept that initiated the discourse, a further line of research also observable in Cluster 1, as heavily interacting with the previous one. Among highly ranked papers we find several approaches regarding the application, or, rather, the problems of application of the biological concept (BSC), mainly in microbiology (Wayne, #3; Dykhuizen, #8; Smith, #11 or, as a case outside microbiology, Knowlton, #20). The association of these topics is well-explained by the fact that the BSC is known as hardly applicable to biological kinds with non-sexual reproduction, such as bacteria and other subjects of microbiology, but also has strange implications to some sexually reproducing kinds as well (e.g. sibling species, Knowlton, #20). What we see in this reference set, then, is best interpreted as a series of responses to the BSC on the part of the practice of systematics.

In sum, Cluster no. 1 can be conceived as quite coherently mirroring what is the bottomline of the XX. century history of the problem, the biological conception (BSC) and the immediate discourse it generated, including both the application and the alternatives of this concept. In terms of the history and philosophy of biology, this profile is the debate over the best theory of species within biology, yielding a theoretically sound category.

- Cluster no. 2: A more recent response: cladistics and the PSC

The core of the second most extensive cluster counts about 100 references (by the knee plot, CDM > 150, as above). Thematically, this group of referred papers is rather coherent. By inspecting the list, striking is the dominance of two concepts, "cladistics" and the "phylogenetic species concept": at least one parameter of each document is related to one of these notions. Many highly ranked references came from the journal *Cladistics*, which has been the main platform of a specific school of systematics by the same name. The reference of the highest rank is Nixon's seminal paper, published in *Cladistics* on the phylogenetic species concept (#1) – this very paper occurred in Cluster no 1 also, but with a relatively low rank, indicating a different emphasis of the two clusters. Papers from other journals also contain "cladistics" and/or a reference to the phylogenetic species concept in their metadata, among their keywords or within their abstracts, with a very few exception. The unity of the profile is also confirmed by the ISI Subject Categories assigned to the papers included: almost each assignment contains "Evolutionary Biology", and, in the majority of the cases, quite exclusively.

Due to this relatively clear profile, Cluster no 2. can be interpreted as the "cladistic response" to the species problem (or, to the BSC). Cladistics is a more recent development in systematics, a school with very specific implications on the definition of the species category, concerning how the phylogenetic tree should be partitioned into species. It is, therefore, closely related to the so-called "phylogenetic species concept" (PSC). The representation of this school is also expressed by the high rank and recurrence of a set of authors, known as the champions of either the phylogenetic or the cladistic conceptualization, e.g. Donoghue, DeQueiroz, Cracraft, Mishler etc. In sum, the cluster is a body of literature on this school of systematics entering the species problem, and producing a significant part of its history.

- Cluster no. 3: the species problem in ecology – a thematic outlier

The core of the *asBC*-cluster no. 3 is a relatively small one, enumerating 15 important references altogether (CDM > 150). Characteristic of its thematic composition are two features of the document set: (1) the references of the two (or three) highest rank are far above the others in terms of weight, and are concerned with the "keystone species concept" (in ecology), and (2) the Subject Category to which these pubs have been assigned by WoS is mainly *Ecology* (and rarely is *Evolutionary Biology*, as opposed to the previous clusters).

This rather compact thematic group is an interesting example of what can be called a "thematic outlier", a strain of research that doesn't belong to the (history of the) very problem under study. Being a "self-contained" group is also reflected in the robustness of the cluster: as noted above, both methods, *cBC* and *asBC* classified these references nearly the same way, as cluster 3 was originated from classical cluster 2 almost without any change (cf. Table 1).

The reason for this sub-topic entering our sample can be said mainly terminological: both discourses are called "species problem" in their own (otherwise, related) contexts. However, while our interest lies in the discourse on the appropriate species concept for biology, the more particular discourse indicated here belongs to the field of ecology and addresses the role of species as actors setting up ecosystems. Therefore, while in the former case the "species problem" stands for the problem of the species concept, in the latter it denotes the problem of finding species in ecosystems (e.g. food webs) whose presence are crucial for its functioning (keystone species). Consequently, in this case, the method (actually, both methods) of bibliographic coupling can be credited for "filtering out" a direction that doesn't belong to the scope of the study.

- Cluster no. 4: An ontology of species taxa for the theory of species

The new cluster no. 4 is also based on a relatively small core, containing about 20 references. The threshold level, CDM > 100, drawn from the knee plot is below the level encountered for the previous clusters, indicating that it is a somewhat less coherent, or more diverse intellectual basis compared to those of the other three groups. A quite interesting multi- (or, as we shall see, rather inter-) disciplinarity can also be observed as to the thematic structure: The pub of the highest rank (ref1) refers to the solution of cladistics to the species problem, yet is has been published in the journal *Biology and Philosophy*, which fact is also reflected in its Subject Category, *History & Philosophy of Science*. This very Subject Category dominates a significant part of the core, together with *Zoology*. What this mixture of "cultures" conveys is a very authentic feature of the species problem, well represented in this separate cluster.

The feature in question is a clear tendency within the XX. Century scientific debate on species to rely on and properly incorporate arguments from the philosophy of science (namely, of biology). Just as Darwin revolutionized systematics by altering the way we look at individual species (species taxa), so did, in the modern history of the problem, two authors, Michael Ghiselin (a biologist) and David Hull (a philosopher of science), the champions of the "individuality thesis" (species as individuals, SAI, Hull 1978, Ghiselin 1974). Addressing the ontology of species (taxa), they argued that species are best viewed, instead of being "classes of organisms", as individuals (particular, historical, evolvable etc. entities). Interestingly, in the technical sense, this view supported some definitions of species, while discrediting others. Among those that could directly rely on SAI was the cladistic species concept and its relatives. As a result of the interaction between biophilosophy and systematics, the SAI and other ontological arguments became integral part of the scientific discourse on species.

This quick historical highlight makes cluster no. 4 a well-interpretable collection. Authors of this cluster are, indeed Ghiselin, Hull and other theoreticians and biophilosophers (Kitcher, Kluge), on one hand, and proponents of the cladistic and phylogenetic concept, on the other (Ridley, DeQuerioz, Mishler, Cracraft etc.). Beyond the symbiotic relation of these two cultures, the presence of the practice of systematics is also present with a high rank (#2). This indicates that theorizing on the status of species propagated into the very circles of practitioners of systematics as well. In sum, cluster 4 can be conceived as a trace of the debate on the ontology of species taxa, being infiltrated into biological theorizing about the species concept (category).

- Cluster no. 1/1 and 1/4: Acquiring historical coherence

The remaining two groups we took under closer inspection were both a fragment of no. 1 described above. The main reason for looking into the internal structure of the first cluster was to sharpen the characterization of how the age-sensitive restructuring of the corpus affected the original thematic groups.

Cluster 1/1 is the fraction of our new cluster no. 1 (The BSC-related theme), that came from the original cluster 1. Recall, that the striking change from the re-clustering procedure was the division of old cluster 1 into new ones, exposed so far as the new cluster no. 1 (considering the majority of its content) and 2. However, it is somewhat more sound to speak of new cluster 1/1 and 2 as the resulting groups. Now, by turning to the content of 1/1, we encounter an even more concentrated profile, than that of the whole class: in this fragment, the position papers proposing and discussing the BSC and its major alternatives exhibit themselves, that is, theorizing of the main figures of biosystematics about the species concept (category). Even more telling, with respect to the capacity of the age-sensitive method, if we compare

the age distribution of references in cluster 1/1 and 2, respectively, that is, between the two descendant of the same old cluster. According to Figure 14, the *asBC* procedure sorted the content of the old cluster into a "more classical", and a "more recent" discussion. For cluster 1/1, references are distributed almost equally before and after the '90s, with a peak in the late 80's, while for cluster 2 the majority of references originate from the '90s, their peak is in the early '90s, and show a more "continuous" or coherent discourse. In other words, the procedure identified the BSC-based dispute (cluster 1/1) as a more classical context, within which the new cluster no. 2, that is, the cladistic/phylogenetic discourse emerged as a more recent movement. Note, that these two, historically distinguishable movements were inseparably linked together by the *cBC* method, in one, thematically coherent but giant cluster. In this sense, the *asBC* method did produce a historically informed thematic structure, differentiating between "ancient" and "new" features of a thematic group.

Considering the contribution of 1/4, the fraction of the BSC-theme that came from the classical cluster no. 4, the picture gets even more interesting. In this small fragment (the core contains only 12 pubs) papers (references) from the very practice of biosystematics are added to the theoretical debate in 1/1, belonging, in particular, to the field of microbiology. This phenomenon recalls our previous observation that new cluster no. 1 covers both (1) the theoretical debate initiated by the biological species concept (BSC) and (2) its extension from, mainly, microbiology, whereby the application of BSC has always been problematic. At this point, we can see that not only does this cluster unify these references, but also "collects" them by "cutting out" the theoretical and the applied part of the BSC-debate from old clusters 1 and 4.

In sum, results suggest that the proposed method of *asBC* has been capable of better identifying strains of research or schools in the modern history of the species problem. On one hand, the *asBC* eliminated a more recent

school within the theoretical discourse, namely, the phylogenetic approach and cladistics emerging from the pool of species concepts. On the other hand, it unified references that show the real or causal, that is, historical unfolding of ideas, instead of reflecting mere topical similarities. This latter feature is shown in connecting the theory and application of the BSC, while, in the original cluster structure these pubs were sorted into the big “theoretical cluster” (old cluster 1), and the “cluster of applications”, mainly, topics in microbiology (old cluster 4), respectively.

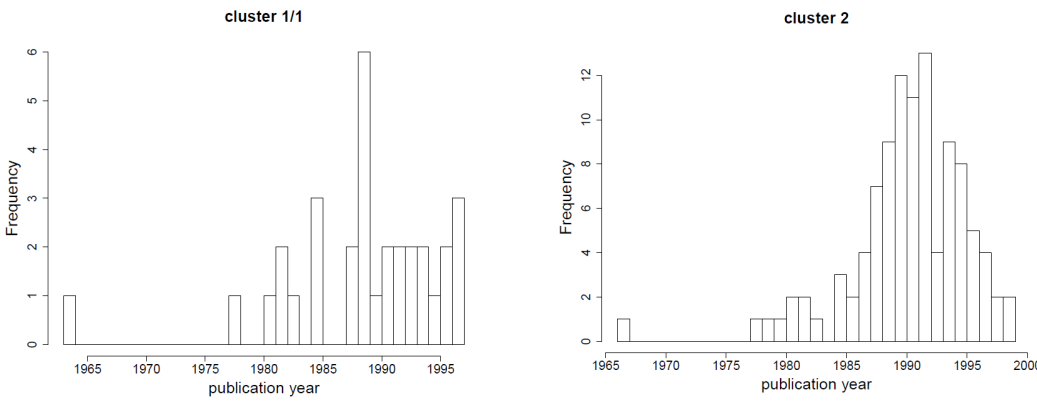


Figure 14. Age distribution of references within the core of clusters 1/1 and 2, respectively.

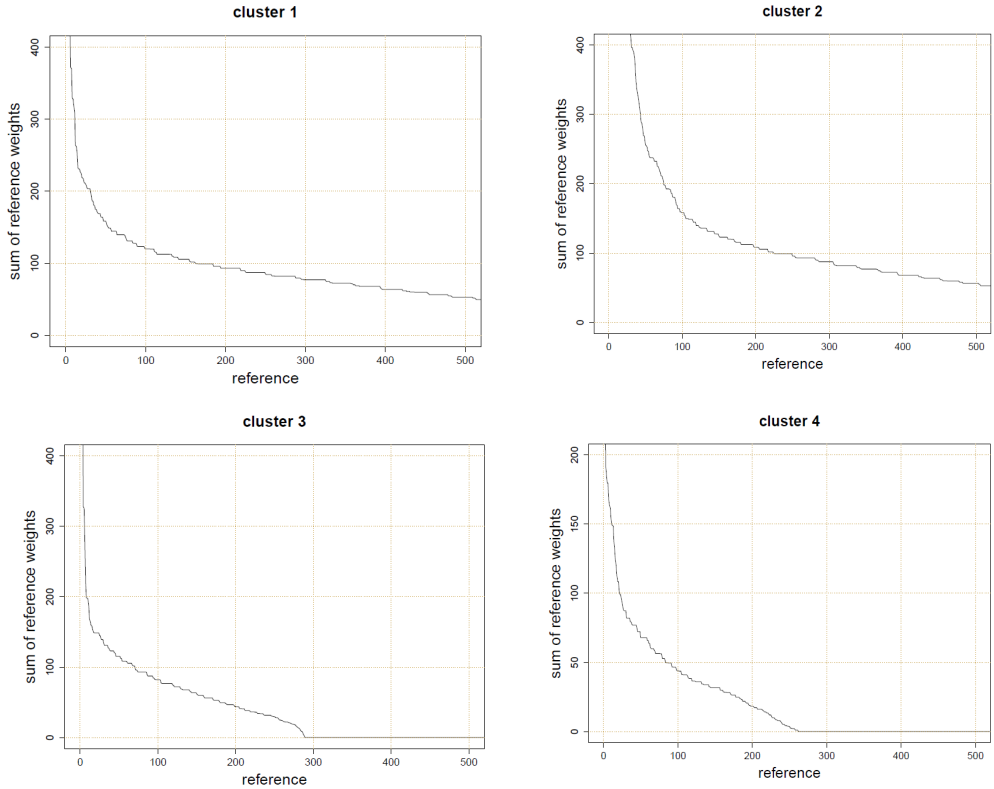


Figure 15. The "knee plots" of clusters 1–4, respectively, supporting the extraction of core references for each. Cumulative weights are plotted against the indices of ranked references. Only the section of the whole curve is graphed where its "knee" is observable.

Table 3. The lists of core references within asBC-clusters 1–4, respectively. In this excerpt, only journal publications are listed. Items are ranked according to their cumulative weight, referred by "Sum of weights"

(age-related weight of reference R *× number of occurrences of reference* R *within the cluster).*

Cluster 1

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
1	MALLET J, 1995, TRENDS ECOL EVOL, V10, P294	644,56	A SPECIES DEFINITION FOR THE MODERN SYNTHESIS	Ecology; Evolutionary Biology; Genetics & Heredity
2	COYNE JA, 1988, SYST ZOOL, V37, P190	446,37	DO WE NEED A NEW SPECIES CONCEPT	Zoology
3	WAYNE LG, 1987, INT J SYST BACTERIOL, V37, P463	329,43	REPORT OF THE AD-HOC-COMMITTEE ON RECONCILIATION OF APPROACHES TO BACTERIAL SYSTEMATICS	Microbiology
4	MAYR E, 1992, AM J BOT, V79, P222	328,07	A LOCAL FLORA AND THE BIOLOGICAL SPECIES CONCEPT	Plant Sciences
5	Mann DG, 1996, HYDROBIOLOGIA, V336, P19	264,15	BIODIVERSITY, BIOGEOGRAPHY AND CONSERVATION OF DIATOMS	Marine & Freshwater Biology
6	VALBONESI A, 1988, J PROTOZOOL, V35, P38	255,07	AN INTEGRATED STUDY OF THE SPECIES PROBLEM IN THE EUPLOTES-CRASSUS-MINUTA-VANNUS GROUP	Zoology
7	COLEMAN AW, 1994, J PHYCOL, V30, P80	232,68	MOLECULAR DELINEATION OF SPECIES AND SYNGENS IN VOLVOCACEAN GREEN-ALGAE (CHLOROPHYTA)	Plant Sciences; Marine & Freshwater Biology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
8	DYKHUIZEN DE, 1991, J BACTERIOL, V173, P7257	231,01	RECOMBINATION IN ESCHERICHIA-COLI AND THE DEFINITION OF BIOLOGICAL SPECIES	Microbiology
9	SMITH JM, 1991, NATURE, V349, P29	231,01	LOCALIZED SEX IN BACTERIA	Multidisciplinary Sciences
10	MISHLER BD, 1982, SYST ZOOL, V31, P491	219,46	SPECIES CONCEPTS - A CASE FOR PLURALISM	Zoology
11	SMITH JM, 1993, P NATL ACAD SCI USA, V90, P4384	218,42	HOW CLONAL ARE BACTERIA	Multidisciplinary Sciences
12	GIANNI A, 1990, EUR J PROTISTOL, V26, P142	216,91	AUTOECOLOGICAL AND MOLECULAR APPROACH TO THE SPECIES PROBLEM IN THE EUPLOTES-VANNUS-CRASSUS-MINUTA GROUP (CILIOPHORA, HYPOTRICHIDA)	Microbiology
13	WILEY EO, 1978, SYST ZOOL, V27, P17	206,02	EVOLUTIONARY SPECIES CONCEPT RECONSIDERED	Zoology
14	MANN DG, 1989, PLANT SYST EVOL, V164, P215	203,69	THE SPECIES CONCEPT IN DIATOMS - EVIDENCE FOR MORPHOLOGICALLY DISTINCT, SYMPATRIC GAMODEMES IN 4 EPIPELIC SPECIES	Plant Sciences; Evolutionary Biology
15	MASTERS JC, 1989, SYST ZOOL, V38, P270	203,69	WHY WE NEED A NEW GENETIC SPECIES CONCEPT	Zoology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
16	SCHLEGEL M, 1988, EUR J PROTISTOL, V24, P22	191,3	TAXONOMY AND PHYLOGENETIC RELATIONSHIP OF 8 SPECIES OF THE GENUS EUPLOTES (HYPOTRICHIDA, CILIOPHORA) AS REVEALED BY ENZYME ELECTROPHORESIS	Microbiology
17	CAPRETTE CL, 1994, J EUKARYOT MICROBIOL, V41, P316	186,15	QUANTITATIVE-ANALYSES OF INTERBREEDING IN POPULATIONS OF VANNUS-MORPHOTYPE EUPLOTES, WITH SPECIAL ATTENTION TO THE NOMINAL SPECIES E-VANNUS AND EUPLOTES-CRASSUS	Microbiology
18	NIXON KC, 1990, CLADISTICS, V6, P211	180,76	AN AMPLIFICATION OF THE PHYLOGENETIC SPECIES CONCEPT	Evolutionary Biology
19	WOESE CR, 1987, MICROBIOL REV, V51, P221	179,69	BACTERIAL EVOLUTION	Microbiology
20	KNOWLTON N, 1993, ANNU REV ECOL SYST, V24, P189	174,73	SIBLING SPECIES IN THE SEA	Ecology; Evolutionary Biology
21	SONNEBORN TM, 1975, T AM MICROSC SOC, V94, P155	171,78	PARAMECIUM-AURELIA COMPLEX OF 14 SIBLING SPECIES	Microscopy
22	FOX GE, 1992, INT J SYST BACTERIOL, V42, P166	164,03	HOW CLOSE IS CLOSE - 16S RIBOSOMAL-RNA SEQUENCE IDENTITY MAY NOT BE SUFFICIENT TO GUARANTEE SPECIES IDENTITY	Microbiology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
23	GRANT PR, 1992, SCIENCE, V256, P193	164,03	HYBRIDIZATION OF BIRD SPECIES	Multidisciplinary Sciences
24	VALBONESI A, 1992, J PROTOZOOL, V39, P45	164,03	THE SPECIES PROBLEM IN A CILIATE WITH A HIGH MULTIPLE MATING TYPE SYSTEM, EUPLOTES-CRASSUS	Zoology
25	BARTON NH, 1985, ANNU REV ECOL SYST, V16, P113	158,6	ANALYSIS OF HYBRID ZONES	Ecology; Evolutionary Biology
26	FELSENSTEIN J, 1985, EVOLUTION, V39, P783	158,6	CONFIDENCE-LIMITS ON PHYLOGENIES - AN APPROACH USING THE BOOTSTRAP	Ecology; Evolutionary Biology; Genetics & Heredity
27	Berlocher SH, 1996, HEREDITY, V77, P83	158,49	POPULATION STRUCTURE OF RHAGOLETIS POMONELLA, THE APPLE MAGGOT FLY	Ecology; Evolutionary Biology; Genetics & Heredity
28	Finlay BJ, 1996, Q REV BIOL, V71, P221	158,49	BIODIVERSITY AT THE MICROBIAL LEVEL: THE NUMBER OF FREE-LIVING CILIATES IN THE BIOSPHERE	Biology
29	MEDLIN LK, 1991, J PHYCOL, V27, P514	154,01	MORPHOLOGICAL AND GENETIC-VARIATION WITHIN THE DIATOM SKELETONEMA-COSTATUM (BACILLARIOPHYTA) - EVIDENCE FOR A NEW SPECIES, SKELETONEMA-PSEUDOCOSTATUM	Plant Sciences; Marine & Freshwater Biology
30	VANVALEN L, 1976, TAXON, V25, P233	152,06	ECOLOGICAL SPECIES, MULTISPECIES, AND OAKS	Plant Sciences; Evolutionary Biology

Cluster 2

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
1	NIXON KC, 1990, CLADISTICS, V6, P211	2205,23	AN AMPLIFICATION OF THE PHYLOGENETIC SPECIES CONCEPT	Evolutionary Biology
2	DAVIS JI, 1992, SYST BIOL, V41, P421	1353,28	POPULATIONS, GENETIC-VARIATION, AND THE DELIMITATION OF PHYLOGENETIC SPECIES	Evolutionary Biology
3	DONOGHUE MJ, 1985, BRYOLOGIST, V88, P172	1321,69	A CRITIQUE OF THE BIOLOGICAL SPECIES CONCEPT AND RECOMMENDATIONS FOR A PHYLOGENETIC ALTERNATIVE	Plant Sciences
4	DEQUEIROZ K, 1988, CLADISTICS, V4, P317	1275,35	PHYLOGENETIC SYSTEMATICS AND THE SPECIES PROBLEM	Evolutionary Biology
5	BAUM DA, 1995, SYST BOT, V20, P560	644,56	CHOOSING AMONG ALTERNATIVE PHYLOGENETIC SPECIES CONCEPTS	Plant Sciences; Evolutionary Biology
6	DEQUEIROZ K, 1990, CLADISTICS, V6, P61	614,57	PHYLOGENETIC SYSTEMATICS OR NELSONS VERSION OF CLADISTICS	Evolutionary Biology
7	WHEELER QD, 1990, CLADISTICS, V6, P77	614,57	ANOTHER WAY OF LOOKING AT THE SPECIES PROBLEM - A REPLY TO DEQUEIROZ AND DONOGHUE	Evolutionary Biology
8	DEQUEIROZ K, 1990, CLADISTICS, V6, P83	578,42	PHYLOGENETIC SYSTEMATICS AND SPECIES REVISITED	Evolutionary Biology
9	MALLET J, 1995, TRENDS ECOL EVOL, V10, P294	545,4	A SPECIES DEFINITION FOR THE MODERN SYNTHESIS	Ecology; Evolutionary Biology; Genetics & Heredity

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
10	DAVIS JI, 1991, SYST BOT, V16, P431	539,02	ISOZYME VARIATION AND SPECIES DELIMITATION IN THE PUCCINELLIA-NUTTALLIANA COMPLEX (POACEAE) - AN APPLICATION OF THE PHYLOGENETIC SPECIES CONCEPT	Plant Sciences; Evolutionary Biology
11	CRACRAFT J, 1992, CLADISTICS, V8, P1	533,11	THE SPECIES OF THE BIRDS-OF-PARADISE (PARADISAEIDAE) - APPLYING THE PHYLOGENETIC SPECIES CONCEPT TO A COMPLEX PATTERN OF DIVERSIFICATION	Evolutionary Biology
12	OHARA RJ, 1993, SYST BIOL, V42, P231	524,2	SYSTEMATIC GENERALIZATION, HISTORICAL FATE, AND THE SPECIES PROBLEM	Evolutionary Biology
13	NELSON G, 1989, CLADISTICS, V5, P275	509,23	CLADISTICS AND EVOLUTIONARY MODELS	Evolutionary Biology
14	BAUM D, 1992, TRENDS ECOL EVOL, V7, P1	492,1	PHYLOGENETIC SPECIES CONCEPTS	Ecology; Evolutionary Biology; Genetics & Heredity
15	DOYLE JJ, 1992, SYST BOT, V17, P144	492,1	GENE TREES AND SPECIES TREES - MOLECULAR SYSTEMATICS AS ONE-CHARACTER TAXONOMY	Plant Sciences; Evolutionary Biology
16	MCKITRICK MC, 1988, CONDOR, V90, P1	478,26	SPECIES CONCEPTS IN ORNITHOLOGY	Ornithology
17	FROST DR, 1990, HERPETOLOGICA, V46, P87	469,97	SPECIES IN CONCEPT AND PRACTICE - HERPETOLOGICAL APPLICATIONS	Zoology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
18	VRANA P, 1992, CLADISTICS, V8, P67	451,09	INDIVIDUAL ORGANISMS AS TERMINAL ENTITIES - LAYING THE SPECIES PROBLEM TO REST	Evolutionary Biology
19	MISHLER BD, 1982, SYST ZOOL, V31, P491	438,91	SPECIES CONCEPTS - A CASE FOR PLURALISM	Zoology
20	DEQUEIROZ K, 1994, TRENDS ECOL EVOL, V9, P27	418,83	TOWARD A PHYLOGENETIC SYSTEM OF BIOLOGICAL NOMENCLATURE	Ecology; Evolutionary Biology; Genetics & Heredity
21	DEQUEIROZ K, 1992, ANNU REV ECOL SYST, V23, P449	410,09	PHYLOGENETIC TAXONOMY	Ecology; Evolutionary Biology
22	DOYLE JJ, 1995, SYST BOT, V20, P574	396,65	THE IRRELEVANCE OF ALLELE TREE TOPOLOGIES FOR SPECIES DELIMITATION, AND A NONTOPOLOGICAL ALTERNATIVE	Plant Sciences; Evolutionary Biology
23	DEQUEIROZ K, 1988, PHILOS SCI, V55, P238	382,6	SYSTEMATICS AND THE DARWINIAN REVOLUTION	History & Philosophy Of Science
24	MORITZ C, 1994, TRENDS ECOL EVOL, V9, P373	325,76	DEFINING EVOLUTIONARILY-SIGNIFICANT-UNITS FOR CONSERVATION	Ecology; Evolutionary Biology; Genetics & Heredity
25	AVISE JC, 1987, ANNU REV ECOL SYST, V18, P489	299,48	INTRASPECIFIC PHYLOGEOGRAPHY - THE MITOCHONDRIAL-DNA BRIDGE BETWEEN POPULATION-GENETICS AND SYSTEMATICS	Ecology; Evolutionary Biology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
26	DEQUEIROZ K, 1990, SYST ZOOL, V39, P307	289,21	PHYLOGENY AS A CENTRAL PRINCIPLE IN TAXONOMY - PHYLOGENETIC DEFINITIONS OF TAXON NAMES	Zoology
27	Avise JC, 1997, P NATL ACAD SCI USA, V94, P7748, DOI 10.1073/ pnas.94.15.7748	281,47	PHYLOGENETICS AND THE ORIGIN OF SPECIES	Multidisciplinary Sciences
28	VANEWRIGHT RI, 1991, BIOL CONSERV, V55, P235	269,51	WHAT TO PROTECT - SYSTEMATICS AND THE AGONY OF CHOICE	Biodiversity Conservation; Ecology; Environmental Sciences
29	VILGALYS R, 1991, MYCOLOGIA, V83, P758	269,51	SPECIATION AND SPECIES CONCEPTS IN THE COLLYBIA-DRYOPHILA COMPLEX	Mycology
30	Taylor JW, 1999, CLIN MICROBIOL REV, V12, P126	255,74	THE EVOLUTIONARY BIOLOGY AND POPULATION GENETICS UNDERLYING FUNGAL STRAIN TYPING	Microbiology
31	PAMILO P, 1988, MOL BIOL EVOL, V5, P568	255,07	RELATIONSHIPS BETWEEN GENE TREES AND SPECIES TREES	Biochemistry & Molecular Biology; Evolutionary Biology; Genetics & Heredity
32	CHASE TE, 1990, MYCOLOGIA, V82, P67	253,06	GENETIC-BASIS OF BIOLOGICAL SPECIES IN HETEROBASIDION-ANNOSUM - MENDELIAN DETERMINANTS	Mycology
33	GRAYBEAL A, 1995, SYST BIOL, V44, P237	247,91	NAMING SPECIES	Evolutionary Biology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
34	DEQUEIROZ K, 1992, BIOL PHILOS, V7, P295	246,05	PHYLOGENETIC DEFINITIONS AND TAXONOMIC PHILOSOPHY	History & Philosophy Of Science
35	Geiser DM, 1998, P NATL ACAD SCI USA, V95, P388, DOI 10.1073/pnas.95.1.388	239,97	CRYPTIC SPECIATION AND RECOMBINATION IN THE AFLATOXIN-PRODUCING FUNGUS ASPERGILLUS FLAVUS	Multidisciplinary Sciences
36	AVISE JC, 1989, EVOLUTION, V43, P1192	237,64	GENE TREES AND ORGANISMAL HISTORIES - A PHYLOGENETIC APPROACH TO POPULATION BIOLOGY	Ecology; Evolutionary Biology; Genetics & Heredity
37	KLUGE AG, 1989, CLADISTICS, V5, P291	237,64	METACLADISTICS	Evolutionary Biology
38	RIDLEY M, 1989, BIOL PHILOS, V4, P1	237,64	THE CLADISTIC SOLUTION TO THE SPECIES PROBLEM	History & Philosophy Of Science
39	FROST DR, 1994, CLADISTICS, V10, P259	232,68	A CONSIDERATION OF EPISTEMOLOGY IN SYSTEMATIC BIOLOGY, WITH SPECIAL REFERENCE TO SPECIES	Evolutionary Biology
40	OHARA RJ, 1994, AM ZOOL, V34, P12	232,68	EVOLUTIONARY HISTORY AND THE SPECIES PROBLEM	Zoology
41	PATTON JL, 1994, SYST BIOL, V43, P11	232,68	PARAPHYLY, POLYPHYLY, AND THE NATURE OF SPECIES BOUNDARIES IN POCKET GOPHERS (GENUS-THOMOMYS)	Evolutionary Biology
42	VOGLER AP, 1994, CONSERV BIOL, V8, P354	232,68	DIAGNOSING UNITS OF CONSERVATION MANAGEMENT	Biodiversity Conservation; Ecology; Environmental Sciences

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
43	Koufopanou V, 1997, P NATL ACAD SCI USA, V94, P5478, DOI 10.1073/ pnas.94.10.5478	225,18	CONCORDANCE OF GENE GENEALOGIES REVEALS REPRODUCTIVE ISOLATION IN THE PATHOGENIC FUNGUS COCCIDIOIDES IMMITIS	Multidisciplinary Sciences
44	BOIDIN J, 1986, MYCOTAXON, V26, P319	225,07	INTERCOMPATIBILITY AND THE SPECIES CONCEPT IN THE SAPROBIC BASIDIOMYCOTINA	Mycology
45	WILEY EO, 1978, SYST ZOOL, V27, P17	223,19	EVOLUTIONARY SPECIES CONCEPT RECONSIDERED	Zoology
46	KORNET DJ, 1993, J THEOR BIOL, V164, P407	218,42	PERMANENT SPLITS AS SPECIATION EVENTS - A FORMAL RECONSTRUCTION OF THE INTERNODAL SPECIES CONCEPT	Biology; Mathematical & Computational Biology
47	VILGALYS R, 1990, J BACTERIOL, V172, P4238	216,91	RAPID GENETIC IDENTIFICATION AND MAPPING OF ENZYMATICALLY AMPLIFIED RIBOSOMAL DNA FROM SEVERAL CRYPTOCOCCUS SPECIES	Microbiology
48	FELSENSTEIN J, 1985, EVOLUTION, V39, P783	211,47	CONFIDENCE-LIMITS ON PHYLOGENIES - AN APPROACH USING THE BOOTSTRAP	Ecology; Evolutionary Biology; Genetics & Heredity
49	Huelsenbeck JP, 1996, TRENDS ECOL EVOL, V11, P152	211,32	COMBINING DATA IN PHYLOGENETIC ANALYSIS	Ecology; Evolutionary Biology; Genetics & Heredity

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
50	MAYR E, 1992, AM J BOT, V79, P222	205,04	A LOCAL FLORA AND THE BIOLOGICAL SPECIES CONCEPT	Plant Sciences
51	LUCKOW M, 1995, SYST BOT, V20, P589	198,33	SPECIES CONCEPTS - ASSUMPTIONS, METHODS, AND APPLICATIONS	Plant Sciences; Evolutionary Biology
52	FARRIS JS, 1991, CLADISTICS, V7, P297	192,51	HENNIG DEFINED PARAPHYLY	Evolutionary Biology
53	HARRISON RG, 1991, ANNU REV ECOL SYST, V22, P281	192,51	MOLECULAR-CHANGES AT SPECIATION	Ecology; Evolutionary Biology
54	Kasuga T, 1999, J CLIN MICROBIOL, V37, P653	191,81	PHYLOGENETIC RELATIONSHIPS OF VARIETIES AND GEOGRAPHICAL GROUPS OF THE HUMAN PATHOGENIC FUNGUS HISTOPLASMA CAPSULATUM DARLING	Microbiology
55	CODDINGTON JA, 1988, CLADISTICS, V4, P3	191,3	CLADISTIC TESTS OF ADAPTATIONAL HYPOTHESES	Evolutionary Biology
56	FARRIS JS, 1994, CLADISTICS, V10, P315	186,15	TESTING SIGNIFICANCE OF INCONGRUENCE	Evolutionary Biology
57	MORITZ C, 1994, MOL ECOL, V3, P401	186,15	APPLICATIONS OF MITOCHONDRIAL-DNA ANALYSIS IN CONSERVATION - A CRITICAL-REVIEW	Biochemistry & Molecular Biology; Ecology; Evolutionary Biology
58	CHASE TE, 1990, MYCOLOGIA, V82, P73	180,76	5 GENES DETERMINING INTERSTERILITY IN HETEROBASIDION-ANNOSUM	Mycology
59	KLUGE AG, 1990, BIOL PHILOS, V5, P417	180,76	SPECIES AS HISTORICAL INDIVIDUALS	History & Philosophy Of Science

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
60	O'Donnell K, 1998, MYCOLOGIA, V90, P465	179,97	MOLECULAR SYSTEMATICS AND PHYLOGEOGRAPHY OF THE GIBBERELLA FUJIKUROI SPECIES COMPLEX	Mycology
61	BAKER CS, 1993, P NATL ACAD SCI USA, V90, P8239	174,73	ABUNDANT MITOCHONDRIAL-DNA VARIATION AND WORLDWIDE POPULATION-STRUCTURE IN HUMPBACK WHALES	Multidisciplinary Sciences
62	CHAPPILL JA, 1989, CLADISTICS, V5, P217	169,74	QUANTITATIVE CHARACTERS IN PHYLOGENETIC ANALYSIS	Evolutionary Biology
63	Burt A, 1997, MOL ECOL, V6, P781, DOI 10.1046/j.1365-294X.1997.00245.x	168,88	MOLECULAR MARKERS REVEAL DIFFERENTIATION AMONG ISOLATES OF COCCIDIOIDES IMMITIS FROM CALIFORNIA, ARIZONA AND TEXAS	Biochemistry & Molecular Biology; Ecology; Evolutionary Biology
64	HILLIS DM, 1992, J HERED, V83, P189	164,03	SIGNAL, NOISE, AND RELIABILITY IN MOLECULAR PHYLOGENETIC ANALYSES	Genetics & Heredity
65	ROJAS M, 1992, CONSERV BIOL, V6, P170	164,03	THE SPECIES PROBLEM AND CONSERVATION - WHAT ARE WE PROTECTING	Biodiversity Conservation; Ecology; Environmental Sciences
66	COYNE JA, 1988, SYST ZOOL, V37, P190	159,42	DO WE NEED A NEW SPECIES CONCEPT	Zoology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
67	OHARA RJ, 1988, SYST ZOOL, V37, P142	159,42	HOMAGE TO CLIO, OR, TOWARD AN HISTORICAL PHILOSOPHY FOR EVOLUTIONARY BIOLOGY	Zoology
68	Burt A, 1996, P NATL ACAD SCI USA, V93, P770	158,49	MOLECULAR MARKERS REVEAL CRYPTIC SEX IN THE HUMAN PATHOGEN COCCIDIOIDES IMMITIS	Multidisciplinary Sciences
69	Legge JT, 1996, CONSERV BIOL, V10, P85	158,49	GENETIC CRITERIA FOR ESTABLISHING EVOLUTIONARILY SIGNIFICANT UNITS IN CRYAN'S BUCKMOTH	Biodiversity Conservation; Ecology; Environmental Sciences
70	STEVENS PF, 1991, SYST BOT, V16, P553	154,01	CHARACTER STATES, MORPHOLOGICAL VARIATION, AND PHYLOGENETIC ANALYSIS - A REVIEW	Plant Sciences; Evolutionary Biology

Cluster 3

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
1	MILLS LS, 1993, BIOSCIENCE, V43, P219	1048,4	THE KEYSTONE-SPECIES CONCEPT IN ECOLOGY AND CONSERVATION	Biology
2	MENGE BA, 1994, ECOL MONOGR, V64, P249	1023,81	THE KEYSTONE SPECIES CONCEPT - VARIATION IN INTERACTION STRENGTH IN A ROCKY INTERTIDAL HABITAT	Ecology
3	Power ME, 1996, BIOSCIENCE, V46, P609	633,95	CHALLENGES IN THE QUEST FOR KEYSTONES	Biology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
4	PAINE RT, 1992, NATURE, V355, P73	328,07	FOOD-WEB ANALYSIS THROUGH FIELD MEASUREMENT OF PER-CAPITA INTERACTION STRENGTH	Multidisciplinary Sciences
5	WOOTTON JT, 1994, ECOLOGY, V75, P151	325,76	PREDICTING DIRECT AND INDIRECT EFFECTS - AN INTEGRATED APPROACH USING EXPERIMENTS AND PATH-ANALYSIS	Ecology
6	WOOTTON JT, 1994, ANNU REV ECOL SYST, V25, P443	279,22	THE NATURE AND CONSEQUENCES OF INDIRECT EFFECTS IN ECOLOGICAL COMMUNITIES	Ecology; Evolutionary Biology
7	WOOTTON JT, 1993, AM NAT, V141, P71	218,42	INDIRECT EFFECTS AND HABITAT USE IN AN INTERTIDAL COMMUNITY - INTERACTION CHAINS AND INTERACTION MODIFICATIONS	Ecology; Evolutionary Biology
8	POWER ME, 1995, TRENDS ECOL EVOL, V10, P182	198,33	THE KEYSTONE COPS MEET IN HILO	Ecology; Evolutionary Biology; Genetics & Heredity
9	TILMAN D, 1994, NATURE, V367, P363	186,15	BIODIVERSITY AND STABILITY IN GRASSLANDS	Multidisciplinary Sciences
10	Wootton JT, 1997, ECOL MONOGR, V67, P45	168,88	ESTIMATES AND TESTS OF PER CAPITA INTERACTION STRENGTH: DIET, ABUNDANCE, AND IMPACT OF INTERTIDALLY FORAGING BIRDS	Ecology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
11	LAWTON JH, 1992, NATURE, V355, P19	164,03	ECOLOGY - FEEBLE LINKS IN FOOD WEBS	Multidisciplinary Sciences
12	YODZIS P, 1988, ECOLOGY, V69, P508	159,42	THE INDETERMINACY OF ECOLOGICAL INTERACTIONS AS PERCEIVED THROUGH PERTURBATION EXPERIMENTS	Ecology
13	Leibold MA, 1996, AM NAT, V147, P784	158,49	A GRAPHICAL MODEL OF KEYSTONE PREDATORS IN FOOD WEBS: TROPHIC REGULATION OF ABUNDANCE, INCIDENCE, AND DIVERSITY PATTERNS IN COMMUNITIES	Ecology; Evolutionary Biology
14	COX PA, 1991, CONSERV BIOL, V5, P448	154,01	FLYING FOXES AS STRONG INTERACTORS IN SOUTH-PACIFIC ISLAND ECOSYSTEMS - A CONSERVATION HYPOTHESIS	Biodiversity Conservation; Ecology; Environmental Sciences

Cluster 4

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
1	RIDLEY M, 1989, BIOL PHILOS, V4, P1	339,48	THE CLADISTIC SOLUTION TO THE SPECIES PROBLEM	History & Philosophy Of Science
2	FROST DR, 1990, HERPETOLOGICA, V46, P87	289,21	SPECIES IN CONCEPT AND PRACTICE - HERPETOLOGICAL APPLICATIONS	Zoology
3	DEQUEIROZ K, 1988, CLADISTICS, V4, P317	191,3	PHYLOGENETIC SYSTEMATICS AND THE SPECIES PROBLEM	Evolutionary Biology
4	HULL DL, 1976, SYST ZOOL, V25, P174	167,26	ARE SPECIES REALLY INDIVIDUALS	Zoology

#	Reference (WoS format)	Sum of weights	Title	WoS Category (Subject Category)
5	SIMONETTA AM, 1992, B ZOOL, V59, P447	164,03	PROBLEMS OF SYSTEMATICS .1. A CRITICAL-EVALUATION OF THE SPECIES PROBLEM AND ITS SIGNIFICANCE IN EVOLUTIONARY BIOLOGY	Zoology
6	HULL DL, 1978, PHILOS SCI, V45, P335	154,52	MATTER OF INDIVIDUALITY	History & Philosophy Of Science
7	KITCHER P, 1984, PHILOS SCI, V51, P308	149,04	SPECIES	History & Philosophy Of Science
8	SIMONETTA AM, 1995, B ZOOL, V62, P37	148,74	SOME REMARKS ON THE INFLUENCE OF HISTORICAL BIAS IN OUR APPROACH TO SYSTEMATICS AND THE SO-CALLED SPECIES PROBLEM	Zoology
9	WILEY EO, 1978, SYST ZOOL, V27, P17	137,35	EVOLUTIONARY SPECIES CONCEPT RECONSIDERED	Zoology
10	SIMONETTA AM, 1993, B ZOOL, V60, P323	131,05	PROBLEMS OF SYSTEMATICS .2. THEORY AND PRACTICE IN PHYLOGENETIC STUDIES AND IN SYSTEMATICS	Zoology
11	KLUGE AG, 1990, BIOL PHILOS, V5, P417	108,45	SPECIES AS HISTORICAL INDIVIDUALS	History & Philosophy Of Science
12	NIXON KC, 1990, CLADISTICS, V6, P211	108,45	AN AMPLIFICATION OF THE PHYLOGENETIC SPECIES CONCEPT	Evolutionary Biology

Results based on model 1.2: Conceptual organization based on citation relations

The community detection on the combined author–keyword citation network resulted in 5 major coherent groups, that is, five major discourses could be identified within the history of the problem. These discourses – modularity classes – are presented below in two, complementry ways: for each identified module the the graph is presented (visualized) in a reduced form, omitting less connected nodes for better readability. At the same time, as to the quantititive version, the most important nodes (authors/keywords) based on their PageRank centrality are plotted in the form of a barchart, characterizing the author group and the conceptual system of the module.

1. The phylogenetic and cladistic theory of the species category.

The most extensive discussion, accounting for the largest module in the graph, may clearly be interpreted as the theoretical debate focusing on a species category defined in terms of phylogenetic criteria and theory. By the reduced graph (Figure 20), two qualified species concepts show itself as organizing the discourse: the *Phylogenetic Species Concept*, and the *Genetic Species Concept.* Even more telling is the structure of the subgraph, as evidenced by both the visualization and the centrality-ranking of authors/concepts depicted in Figure 16. The upper part of of the graph (*Ereshefsky, M*, *definition*, *clade*, etc.) mirrors the contribution of philosophers of science and theoreticians of biology to concept formation: the concentration of these approaches is rather striking in the full network of this module (Figure 21, framed area), whereby most influential "philosophers" of the problem are present (*Ghiselin, Hull, Wiley, Sober, Mishler, DeQuiroz, Platnick, Cracraft* etc.), along with a set of thematically related key concepts on the ontology of species (*individual, class, definition, ostensive definition, name*). This group is connected, through a set of central concepts (including concepts from experimental science, such as *mithocondrial DNA*, *DNA barcoding*) to an extended group of approaches addressing species within experimental/molecular biology.

The so-called *genetic species concept* is positioned in this context, while the *phylogenetic species concept*, as such, is positioned in the neighborhood of theoreticians. This configuration of the network corroborates, on one hand, (1) the substantial – interdisciplinary – interaction between the philosophy of science and species systematics. The famous *individuality thesis*, stating that species are ontological individuals instead of classes, is, indeed, seems to penetrate the discussion on the species category, serving as the philosophical background for the *phylogenetic* and – as a highly related definition – the cladistic concept. This interdisciplinarity is also made apparent by the centrality ranking of network members: the high end of the distribution shows *Hull, DL*, the philosopher co-inventor of the individuality thesis along with *mithocondrial DNA* as the third and second most central actor in the net, respectively. It is also of great interest that the tradition of theorizing on the species category, the majority of "philosophers" of the issue, show up almost exclusively in this subdiscourse, that is, in relation to the phylogenetic conception. On the other hand (2), a further important historical connection emerges from this module, between a theoretical and an experimental tradition. Based on the network structure outlined above, it can be hypothesized that the genetic species concept is a a descendant of the phylogenetic species concept, the former being an operationalized or, at least, more applicable version of the latter in the context of experimental, namely molecular, biology.

2. Research on phylogenetic inference

The next module in the list, in terms of graph size, is a well-interpretable and highly coherent research tradition overlapping with the quest for a valid species category. The reduced graph (Figure 22) reveals the discourse on *phylogenetic inference*, that is, the methodology on experimentally inferring and reconstructing phylogenies of/among taxa, including species. *Phyologenetic inference* is both a methodological and experimental subject

within evolutionary biology, as is clearly reflected in the set of constitutent concepts. The structure of the module is indicative of both its relation to the species problem, and of its coherence: through central concepts (*phylogeny, species delimitation, molecular systematics, molecular phylogeny*) two cohesive groups are connected to each other: the set of authors interacting on this subject, and the related conceptual system as an apparent description of the methodological issues involved. The interface of this tradition with the species problem is the valid procedure of experimentally delimiting species (by phylogenetic reconstruction): most interestingly, this experimental methodology applies multiple theoretical species concepts (as evidenced by the nodes *morphological species concept*, *biological species, phylogenetic species*) for the purposes of operationalization. This methodological character is also evidenced by the Page Rank centrality ranking (Figure 17), whereby *parsimony*, as the main axiom or object function of the inference method is shown as far the most central concept, along with the author *Felsenstein*, known for the firs phylogenetic inference software package. Even the long tail of the centrality distribution almost uniformly covers mathematical and experimental methods (*weighted/unweighted least squares, maximum likelihood, Bayesian estimation* etc.). Though not apparent either on the reduced layout or in the ranking plot, by a relatively weak link, Ernst Mayr, the classic figure of the species problem originally proposing the biological species concept, is also classified together with this module. The connection is established through the concept *natural system* (present in the reduced graph), a Darwinian principle rediscovered by Mayr for systematic biology, and – apparently – entertained by this tradition as the primary criterion for selecting among alternative inference methods.

3. Speciation and the BSC tradition

The context one would expect Mayr to appear within would be the next most significant subdiscoruse, which altogether can be referred to as the tradition induced by the Biological Species Concept (BSC). The subgraph, again, mirrors a highly cohesive group (Figure 24): a densely connected set of concepts is being related to a set of interacting authors through, basically, three central terms: *speciation, reproductive isolation* and *hybridization.* Almost all concepts are clearly related to an aspect of the debate on the biological, or interbreeding-based definition of the species category: e.g. *sexual isolation, hybrid inviability, hybrid sterility, gene flow, ring species.* The same phenomenon is being shown via the centrality ranking plot (Figure 18): each constituent in the list of terms is related either to a classic feature of the BSC, or to the original arguments in support of the conception. The most central term, however, is *speciation,* the mechanism of species formation which, with the Biological Concept, took a definitive role as a phenomenon that any theoretically sound species concept should explain. In sum, a natural interpretation of this module is that it is organized by the debate on speciation as framed by the BSC, with all its empirical difficulties caused by the primary criterion of reproductive compatibility/isolation. Having both field science (*vocalizations, sunflower, allozymes*) and theory engaged in the same tradition, the graph also shows the related philosophical influence on the debate: the top part of the reduced network contains *essentialism, natural kinds, levels of selection* in the neighborhood of the *biological species concept*, with related theoreticians (*Hey, Wilkins*). Though the biological conception is often communicated by historians/philosophers of science as the "death of essentialism" (whereby species taxa are no longer natural kinds), these ontological arguments are usually linked to the whole modern history of the species debate: the present result, however, bounds the context of (explicit and terminologically detectable) anti-essentialism more closely to the BSC tradition, which is an additional piece in the historical mosaic of the species problem.

4. Challenges for species concepts on the part of microbiology

The next two modules, though significant in size individually, are best described in a parallel manner, the reason being both represent the same type of contribution to the species problem. As witnessed by historians of biology, theoretically grounded and general species concepts have often been challenged from within different fields of application or the practice of systematics. Especially resistant to definitional approaches is the field of microbiology, as for example in the realm of microorganisms – mostly lacking sexual reproduction – the biological species concept, as such, can hardly work. The two subdiscourses in question cover a related research subject in microbiology, respectively, each of which poses a challenge for theoretical definitions of the species category. Both modules, therefore, convey the reception of the theoretical debate in experimental science. The more extensive (Figure 23) is held together by the central concepts *recombination, evolution, species concept, lateral gene transfer*, which is also confirmed by the centrality ranking (Figure 17), complementing the list with *linkage disequilibrium, bacteria.* The microbiological character of this discourse is reflected in that most constituent terms (the author interaction part aside) are names of microbial taxa. This structure is a good characterization of a quest for a microbial species concept based on phenomena among microorganisms (mostly bacteria) that are comparable to theoretical species critera (as e.g. "recombination through lateral gene transfer"). Even more specific is the other module categorized under these approaches, concerned with a certain taxonomic group called *Diatoms* (Figure 25). Diatoms are a type of phytoplankton or algae, that is also hard to reconcile with existing species definitions. The corresponding subgraph exhibits a set of methods from cell and molecular biology aimed at the task of species delimitation.

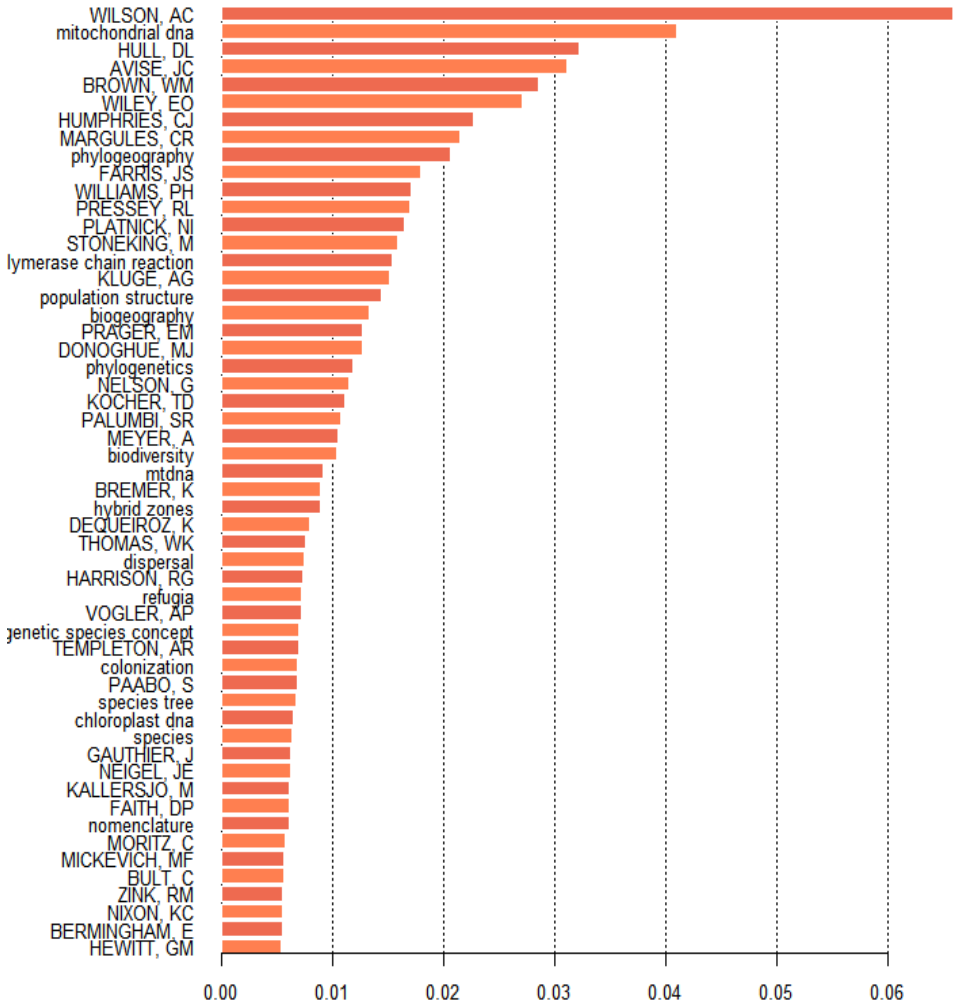


Figure 16. The phylogenetics module

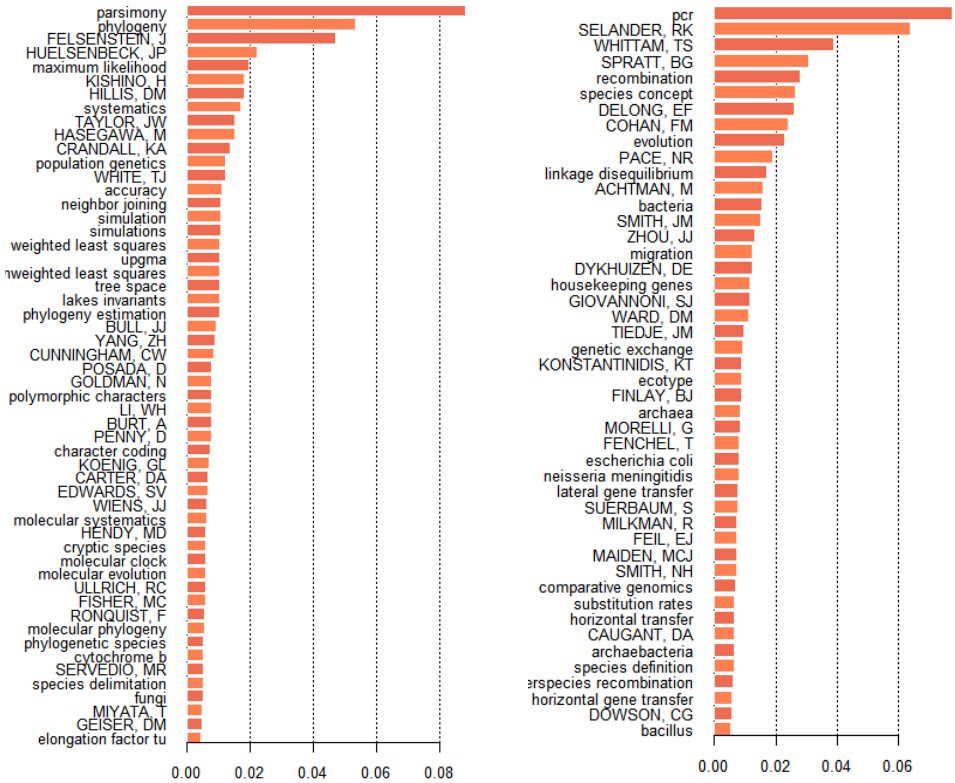


Figure 17. The phylogenetic inference and recombination module

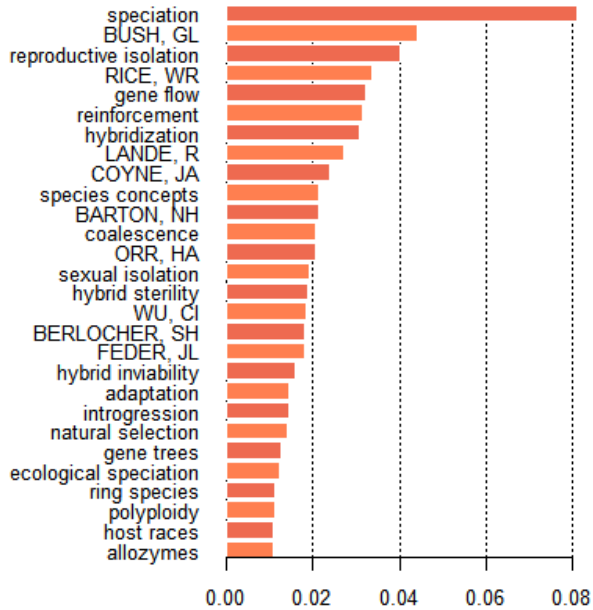


Figure 18. The speciation/BSC module

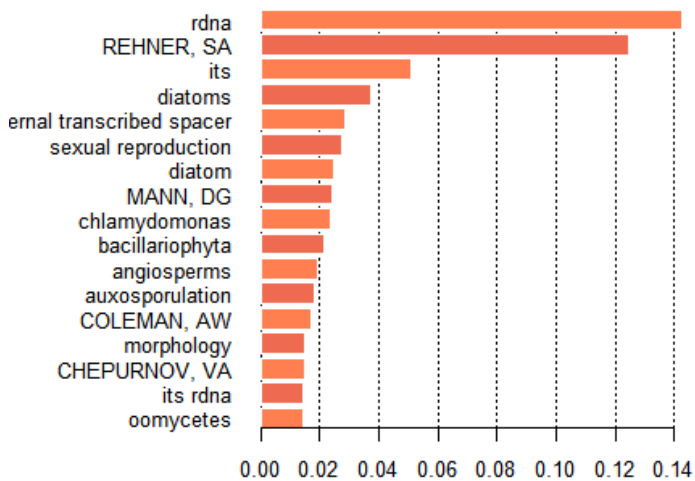


Figure 19. The diatoms module

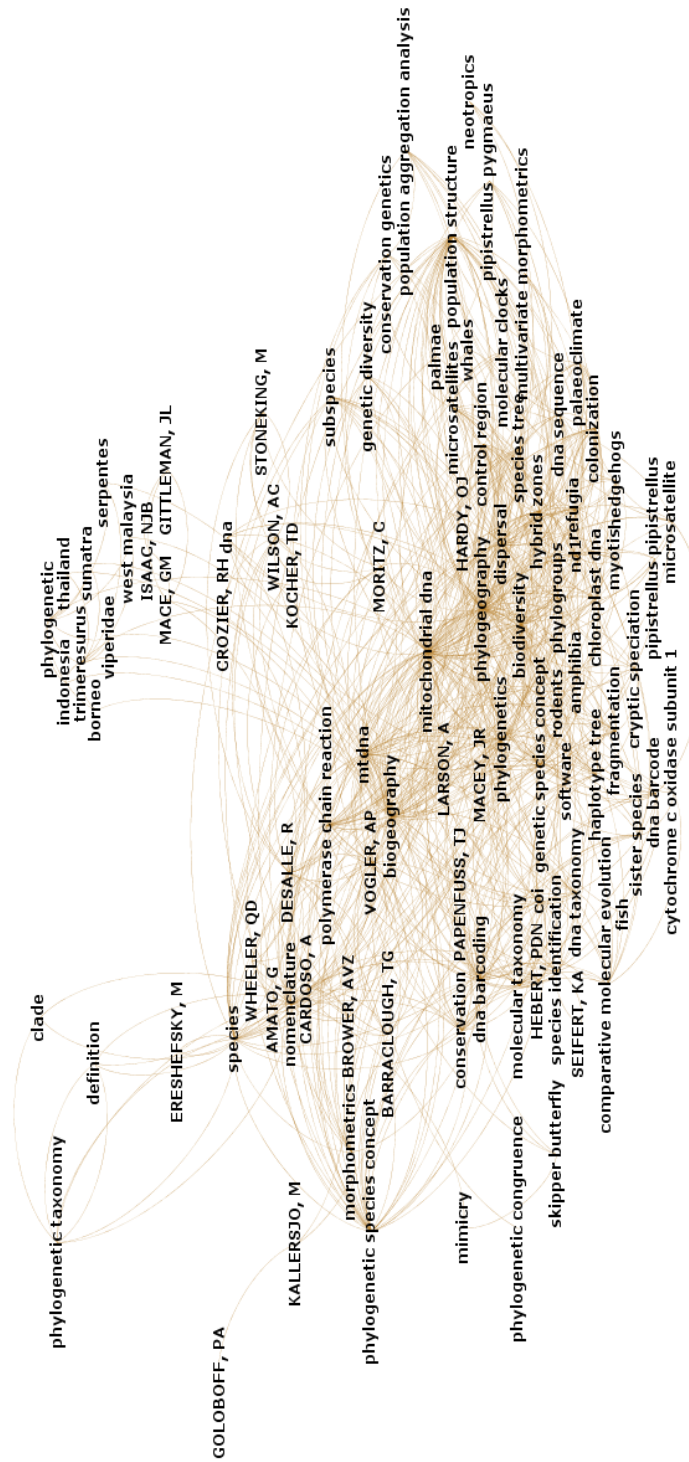


Figure 20. The phylogenetics subgraph

Figure 21. The phylogenetics subgraph, full version

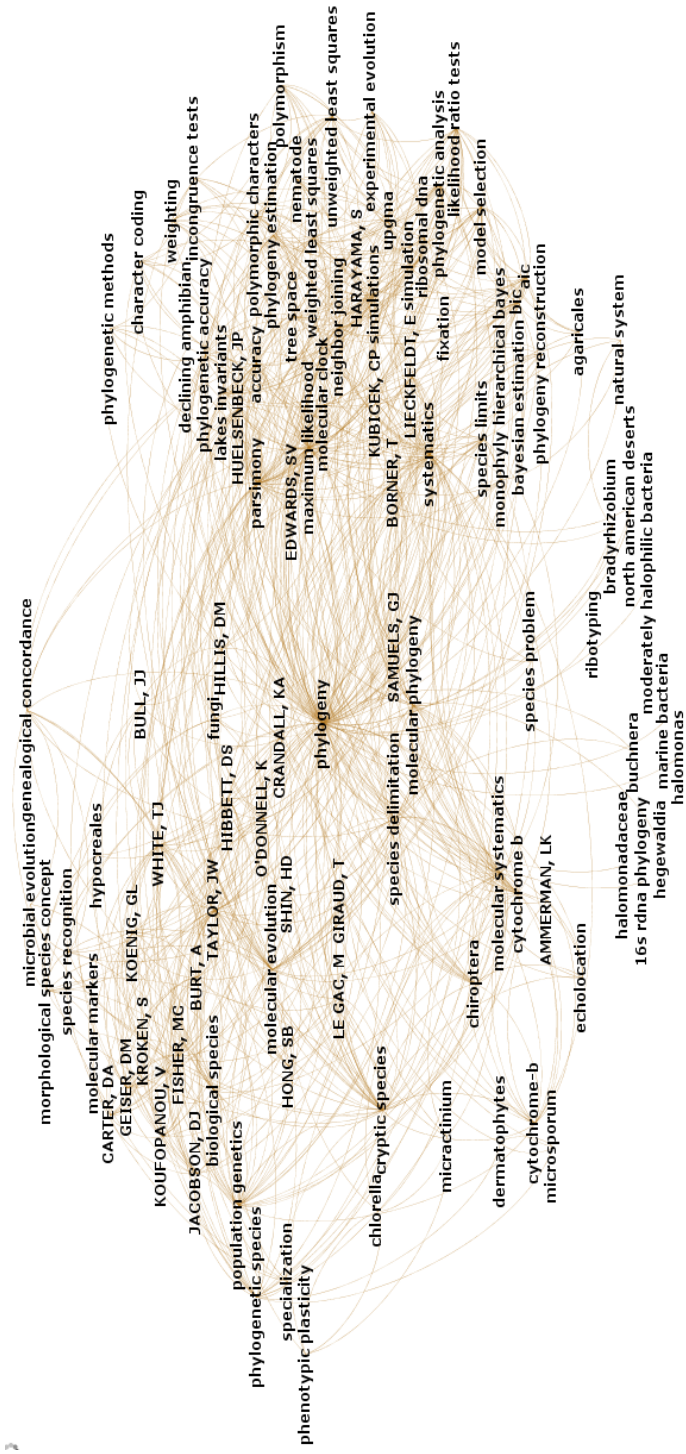


Figure 22. The phylogenetic inference subgraph

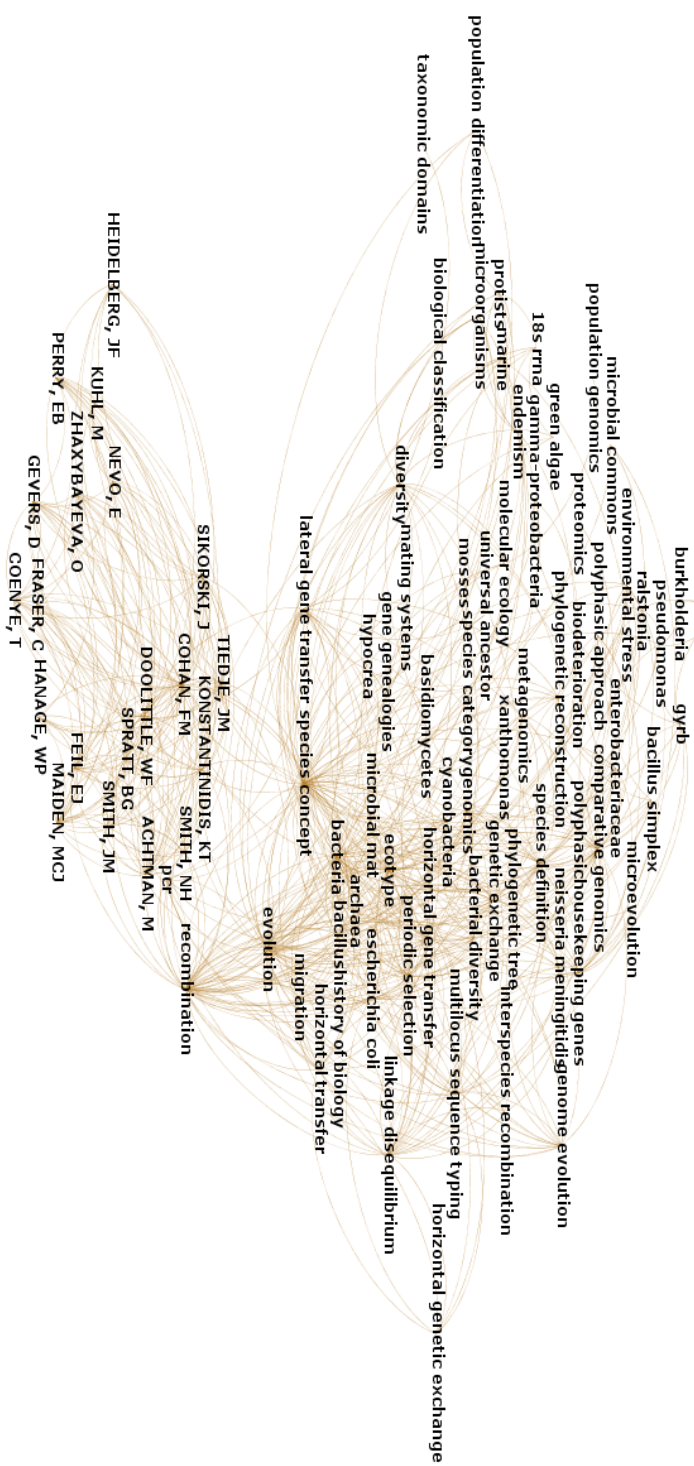


Figure 23. The recombination subgraph

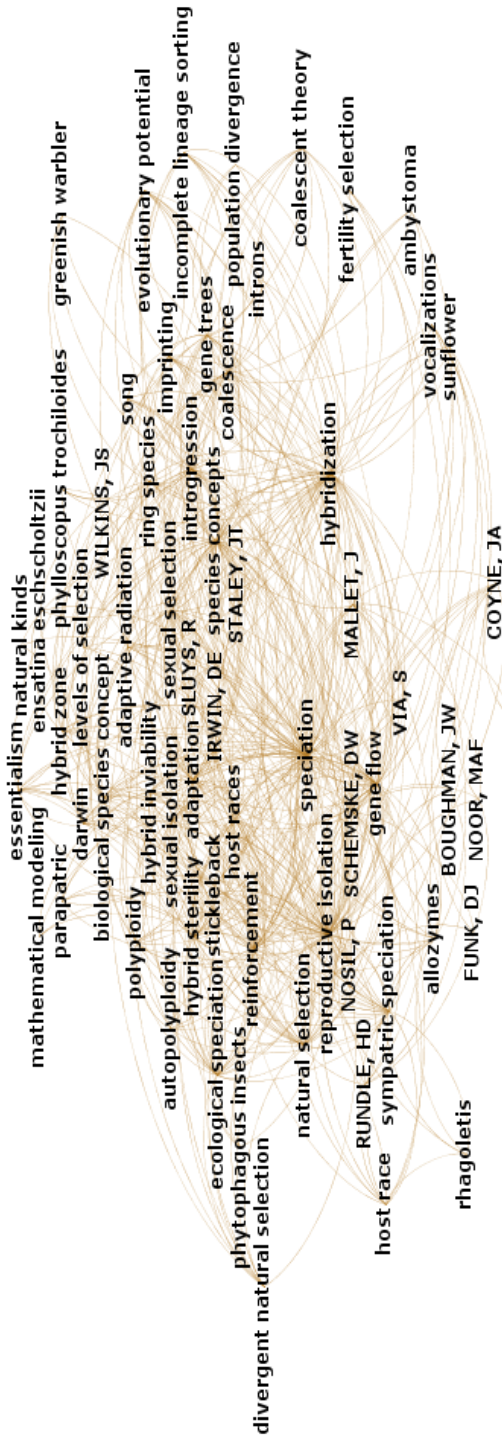


Figure 24. The speciation/BSC subgraph

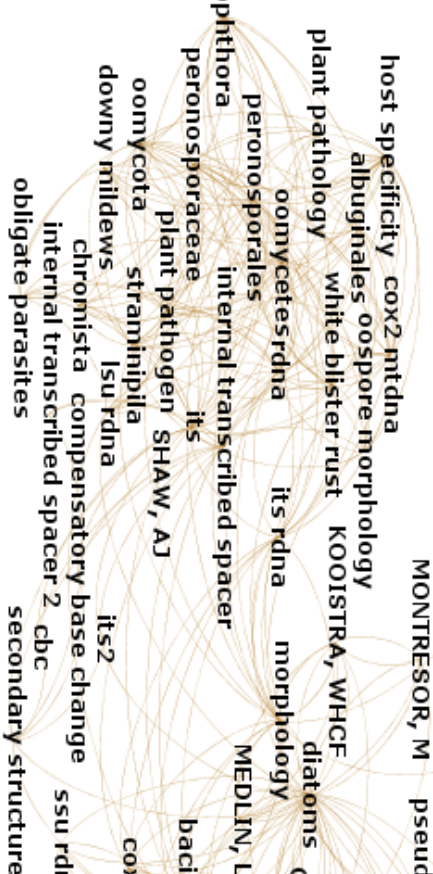


Figure 25. The diatoms subgraph

Results based on model 2: The evolution of the latent conceptual organization

By fully exploiting the analytic capabilities of the TOM framework we can gain a deep insight into the organization of the Species Problem, including the development of the latent conceptual structure of the discourse. To that end, we have devised a TOM-based profile for each topic cluster resulting from the clustering exercise above, composed of the following analytics:

- *The topic overlay map of the cluster.* As the central feature of the proposed methodology, the topic overlay map (TOM) is shown for each cluster. TOMs are depicted via the topic map (basemap) of the whole discourse (Species Problem) with the size of topics (nodes) being proportional to their relative share within the cluster (customized basemap). Consequently, TOMs make the dominant topics and their network, that underlie particular clusters, visible, so that the cognitive basis of the cluster would become transparent. Furthermore, since overlay maps characterize each document cluster in the context of the whole discourse (on the basemap), the internal position and relations of the cluster within the cognitive organization of the Species Problem is also made visible.

- *The keyword profile of the cluster.* As for the comparisons between TOM- and VSM-clusters reported above, a content profile was also generated for document clusters in terms of their most frequent keywords (frequency distribution). The keyword profile is based on author keywords and keywords generated from the titles of document reference lists.

- *The development of relative cluster size along the timeline.* A highly informative view on the evolution of the problem can be obtained by tracking the development of each cluster along the timeline. From this perspective, the internal trends of constituent research directions, the emergence, "rises and falls" of subdiscourses (encoded in document clusters) can be followed and compared, through which a dynamic and

historical picture of cluster structure may be provided. Accordingly, an analytics is provided for each cluster combining three time series: (1) the annual – relative – size of the cluster, that is, the annual % from the whole set of documents comprising the cluster, (2) an 5-year moving average of annual cluster size and, for the purposes of comparison (3) an 5-year moving average on the annual relative size of the *whole* corpus, as the % of docs for each year. In verbal terms, (1)–(2) conveys how the cluster unfolds within the history of the problem, while (3) shows how the whole problem is progressing in parallel. In this way, by comparing the progress of the whole discourse with that of the particular research direction within, the relevance of sub-discourses can be traced and historically located for the SP.

Cluster profiles built in this fashion are collected at the end of this chapter. Striking even from the first and quick overview of these analytics is a general feature of the cognitive–topical organization of the Species Problem, that is clearly represented by the profiles. Most document clusters are based on a combination of the same "central" topic, depicted by node no. 1, and one or two related topic(s), that seem rather specific to the cluster in question. Topic 1 is made up of the issues and concepts most central to the Species Problem in general, while each cluster-specific topic, as we shall see below, covers a well-recognizable context for this "core" (i.e. Topic 1). In other words, all clusters appear to share the core topic (which contributes to all at a variable rate), but are still distinctively characterized by a different perspective (field, theoretical context, context of application etc.). In the following, we organize the overview of profiles according this structural feature of the discourse, that is, by building a typology of the contexts for Topic 1, as evidenced by the overlay maps for individual clusters.

Species problem(s) related to Ecology

The most clearly distinguishable context (or, rather, family of contexts) of the Species Problem is outlined by Clusters 5, 6, and 13. Cluster 5 is dominated by a topic from the study and (mathematical) modelling of ecosystems (topic #9), whereby research focuses on the role of species (as building blocks) in the functioning of such systems (*keystone species concept, equations, dynamics, food webs, Lotka–Volterra system* etc.). Clusters 6 and 13 are marked by topic #11, which represents a related, but conceptually different ecological/environmental direction, *viz*. conservation biology. In this context, the species problem is interpreted as establishing measures of biodiversity (cf. key terms as *biodiversity, richness, umbrella species*), which largely depends on the recognition (and individuation) of species taxa. Therefore, these two "conservation"-clusters are more central to the Species Problem, as introduced here, than the previous "ecosystems"-cluster, which fact is also indicated by the respective overlay maps in two respects. On one hand, the "conservation"-clusters show a dual dominance – interaction – of both the core topic (Topic 1) and the specific context (Topic 11), while the "ecosystems"-cluster is characterized by the specific context alone (Topic 9), with the core topic – the central concepts and issues of the Species Problem – being much more suppressed. On the other hand, the topology of the topic map (basemap) is also indicative in itself, since the "conservation"-topic #5 is part of a densely related topic-group around the core topic #1, while the "ecosystems"-topic is much less connected to this group, being attached to the topic map only through the "conservation"-topic. The relation between the latter two, however, is strong: it can be said that the "ecosystems"-topic is related to the core problem with the mediation of the "conservation"-topic. A further sign of the differing relevance of the ecology-based clusters is exhibited by the timeline-diagramms. Compared to the overall development of the Species Problem (indicated by the red curve with data points), the trendline for Cluster 5 is following a different course, showing an earlier peak (end of the 90's) and a moderate decline afterwards. On the contrary,

both Cluster 6 and 13 follows the main trend more closely, the curves progressing along the overall, ascending trendline. This altogether shows that the "conservation"-clusters are being much more integrated with the main discourse on the Species Problem, while the "ecosystems"-cluster seems to be a separate but overlapping direction. To put it another way, one can detect the combination of different "species problems" here, that are still separable via the TOM-profiles.

The Species Category and the ontology of species

As to the nature of the Species Problem, the most representative clusters, namely Cluster 8, 14 and 15, are characterized by the interplay between the core topic and Topic #5. This latter topic is a collection of issues and concepts related to the ontological status of species taxa (*natural kind, individuals, essentialism, Darwin*) that of the species category (*pluralism, realism*), and, quite tellingly, terms related to the school of systematics called "cladistics" (*cladist, cladogram, german, Hennig*). The cluster profiles clearly show how the ontological – philosophical – issues penetrate into the mainstream biology-based discourse on the definition of species. In cluster 8, an equal weight is given to the core topic and the "ontological" topic, showing how the cladistic approach to systematics and the species concepts gains support from the thesis that species are ontological individuals (*species-as-individuals* thesis, SAI: cf. term frequencies *systematics, individuality, individuals, cladistics, monophyly, names, german*). A similar behavior can be attributed to cluster 14, with more weight given to "biology" (Topic #1). A further difference is that cluster 14 is expressed earlier in the timeline with a peak, preceeding (and anticipating) the rise of this subdiscourse (later also following an ascending timeline in itself). As opposed to these two, cluster 15 is marked by a clear dominance of the "ontology"-cluster over the core topic, which is also mirrored by its textual profile. This cluster coveys the

theoretical (or meta-theoretical and philosophy-rooted) subdiscourse on the ontology of species and its implications on the species concept/category (*classification, natural kinds, individuals, evolution, pluralism, history* being the leading keywords).

The evidence for the proper interaction of philosophy and biology, wittnessed primarily by these profiles, also comes from both the overlay maps and the topology of the underlying topic map. The clear co-activity of the core topic and the "ontology"-topic is, though at a varying rate, universal for these profiles. Even more interestingly, the "ontology"-topic is not part of the dense topic group around the core in terms of network topology, just as the "ecosystems"-topic. Rather, ontological issues relate directly to the core, which links those issues to this central topic group. This picture also confirms the peculiarity of the situation, that an "outsider" discipline – philosophy – directly affects a scientific discourse, otherwise embedded in complex biological context. Furthermore, this arrangements can be taken as evidence not only for the interplay of distant topics, but, on top of that, for proper interdisciplinarity exhibited by the subject matter.

A further evidence for this deep embeddedness of the ontological perspective within the discourse can be drawn from the timeline diagramms. It can be seen for basically all three clusters that their progress goes "hand-in-hand" with the main trendline, reporting a shared dynamics of the general problem and the biophilosophical debate. The profiles alltogether well support the hypothesis behind the factors of the modern SP: in particular, it is made visible how the individuality thesis affects the success of the so-called cladistics-based (and, in derived forms, phylogenetic and genetic) definition of species.

The species problem and specialities of bioscience

Many profiles resulting from this mapping can be grouped into a well-defined family of clusters (or subdiscourses). The feature that collects these profiles together is that each cluster is concentrated around the realization of the species problem within a speciality of bioscience – the latter usually focusing on a particular, but broader taxonomic group (such as fungi or algae). A common characteristics for most of these specialities is that standard species concepts (in fact, the notion of species) is problematic for their purposes, due to the special nature of the subject matter (taxa of interest). Cluster 9, with Topic #12, mostly represents the problems micology in defining species out of fungi; Cluster 11 and Topic #14 stands for the species problem in botanics, plants often exhibiting "irregular" speciation behavior and patterns against the Biological Species Concept (*reproductive isolation, hybridization, speciation, polyploidy, pollination* etc.); Cluster 16 with Topic #14 places the SP in the context of paleontology, whereby the reconstruction of species from the fossil record makes it hard to apply modern definitions (such as the Biological SP) based on "observable" relations between existing taxa (reproductive isolation). A deeply interlinked subgroup of clusters is concerned with microbiological taxonomy: Topic #3 and Topic #2 underlies a couple of profiles related to the taxonomy of algae, more specifically, diatoms, whereby it is extremely problematic to empirically systematize biological diversity (Cluster 17, Cluster 20). Cluster 18 is specifically focuses on systematic bacteriology, bacterial phenomena escaping most approaches to defining species in this realm, represented by Topic #0.

This collection of cluster profiles also exhibits an important structural aspect of the Species Problem. As can be seen from the outlines above, each cluster is a combination of the core topic and a cluster-specific topic. According to the overlays, each of these cluster-specific nodes belongs to the dense central topic group directly surrounding the core (Topic #1), in terms of network topology. That is to say, these clusters provide the main context in which

the SP exits and develops. This structural feature provides further evidence that the Species Problem has been, and is being mainly fed by the problems of applying theoretical concepts in the practice of biology ("field research"). To put it another way, this arrangements shed light on a main dimension of the SP, viz. the difficulties to achieve the theory-driven goal of formulating a species concept that is universal enough to cover the diversity of the living world. The timeline diagramms also support this interpretation by indicating a high fit between the main trendline and the respective cluster curves.

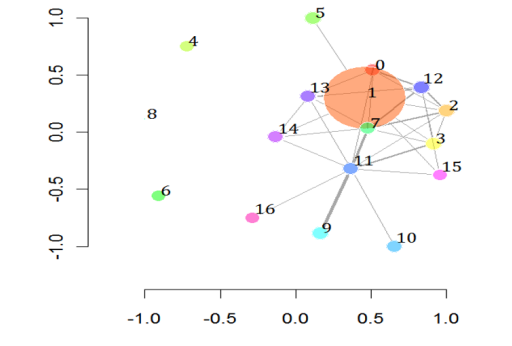
The Species Problem and methods from molecular biology

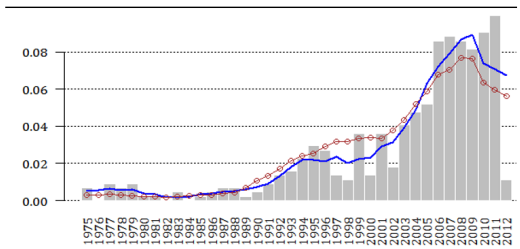
Still part of both the speciality-driven clusters is Cluster 10 focusing on the delimitation of species taxa with the aim of molecular biology. The distinctive topic (Topic #2) unerlying this cluster is also part of the central group of basemap nodes. The reason for still treating this profile separately is twofold: instead of specific taxa, this subdiscourse is concerned with "methodological paradigm" that cross-fertilizes taxonomic schools and different approaches to the species concept, and, nevertheless, represents the present "instrumentalist" consensus on the species problem. Relying on genetic and molecular markers (cf. *RNA secondary structure, ribosomal RNA, mtDNA, molecular phylogeny, polymerase chain reaction*) to separate species by inferred phylogenies is a pragmatic approach that accomodates features from many theoretical species concepts (Phylogenetic, Cladistic, Genetic, Biological), while practically overlooking conceptual problems. This cluster can, therefore, be seen as the response of normal science to the theoretical debate with an "inference to the best explanation"-type framework.

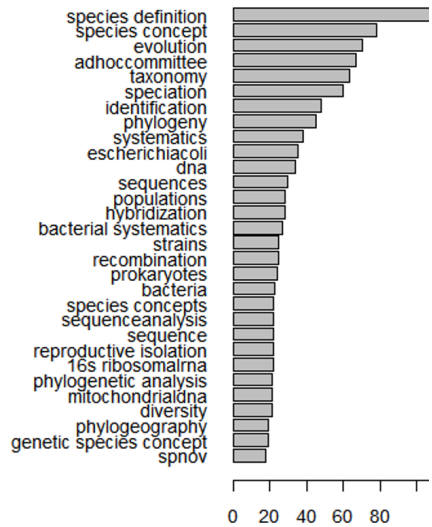
The Species Problem and evolutionary mechanisms. Somewhat distributed or scatterred among clusters is a "horizontal" theme that can still be well recognized from browsing the profiles. This theme covers the research in evolutionary biology concerned with *speciation* and *speciation mechanisms*.

This issue is inherently related (sometimes even hardly separable) to the definition of the species category. It is widely assumed that had the “natural” mechanism(s) that isolates species been found, a corresponding definition based on this/these mechanisms would also naturally follow. Cluster 12 heavily relies on this theme (*sympatric speciation, adaptive radiation, reproductive isolation, hybridization*), though with a focus on viruses with Topic #13 (belonging, in that sese, to the taxa-specific core group). It is also well-recognizable in Cluster 11 with Topic #14 (*reproductive isolation, hybridization*), which similarly furhter specialzes in the study of a specific taxon (plants).

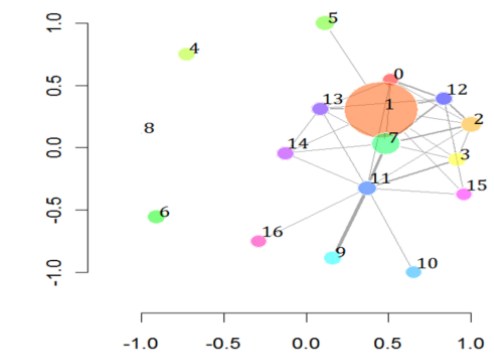
Cluster 1

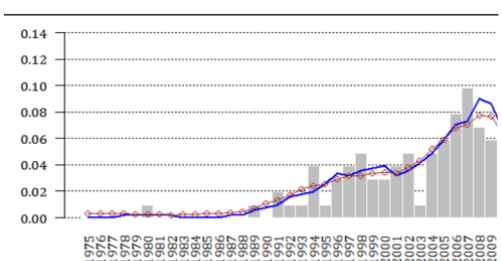


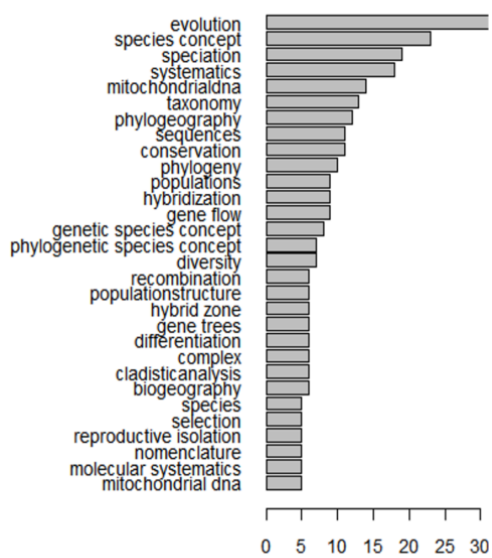




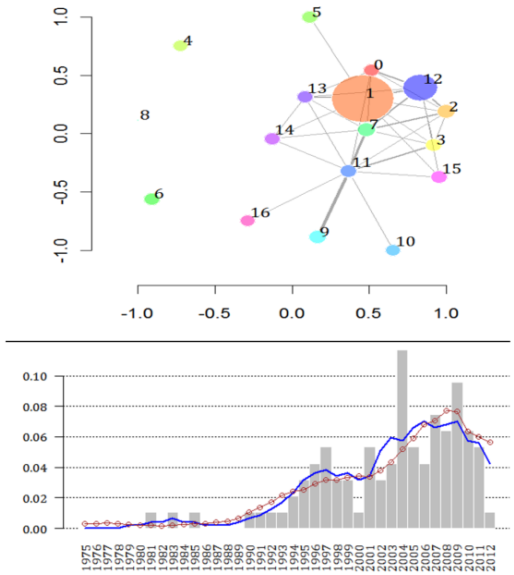
Cluster 2

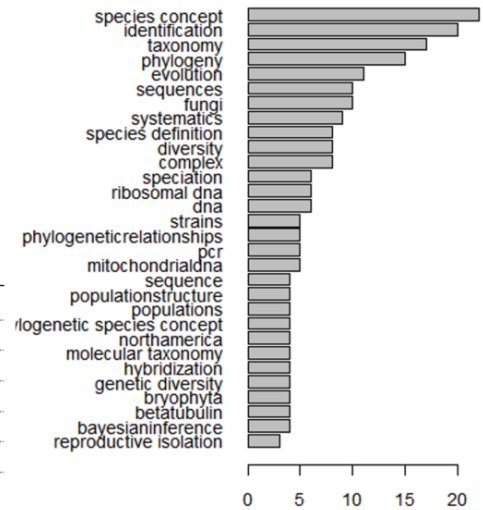




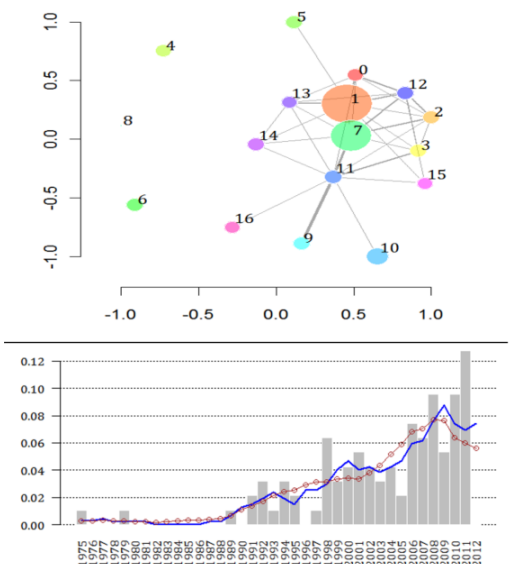


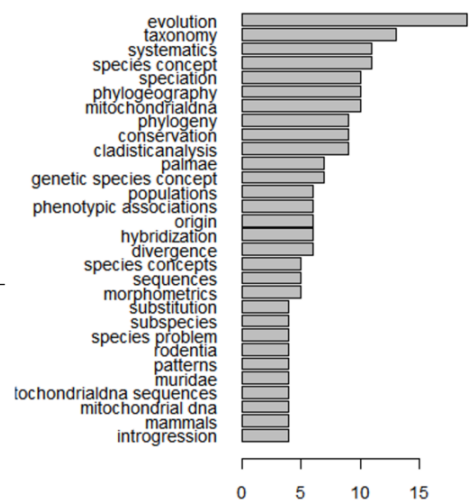
Cluster 3



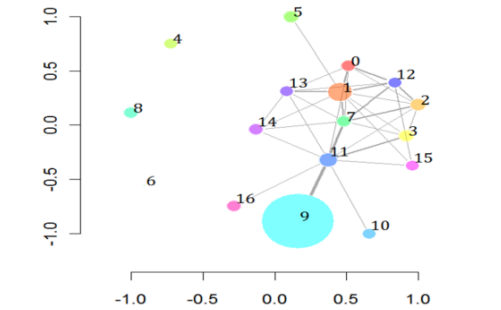


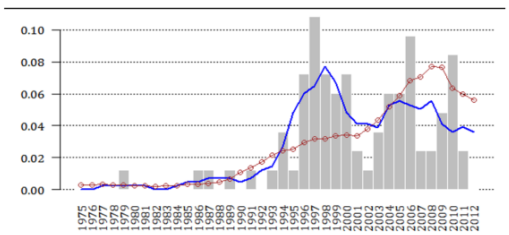
Cluster 4

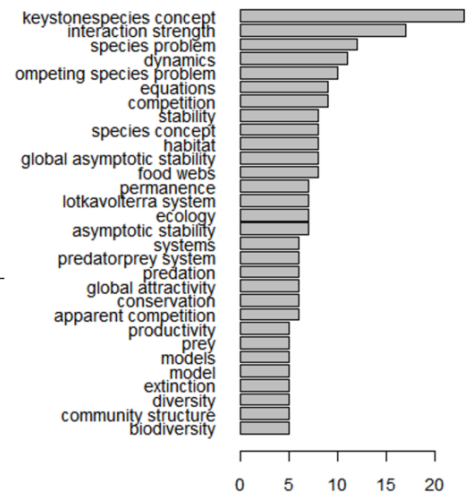




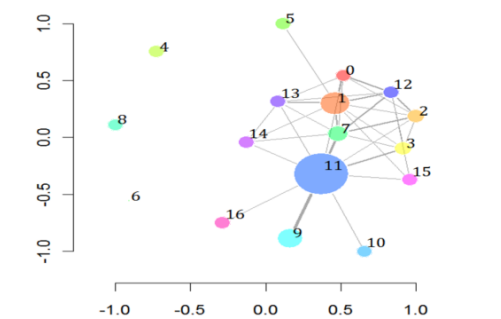
Cluster 5

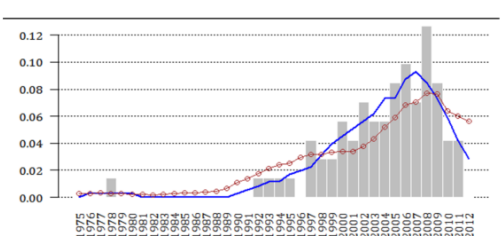


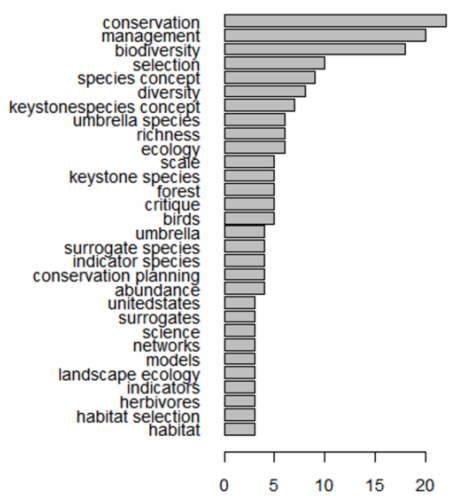




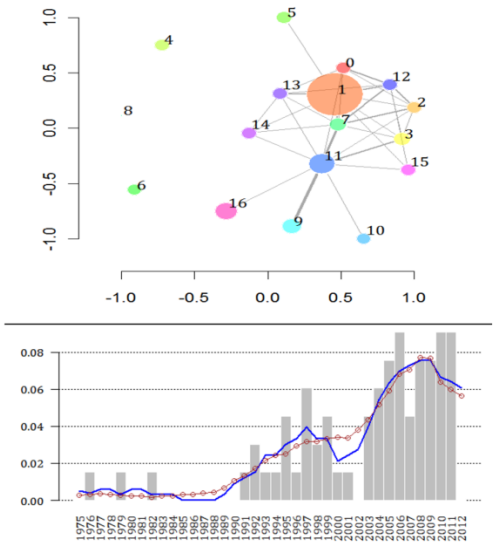
Cluster 6

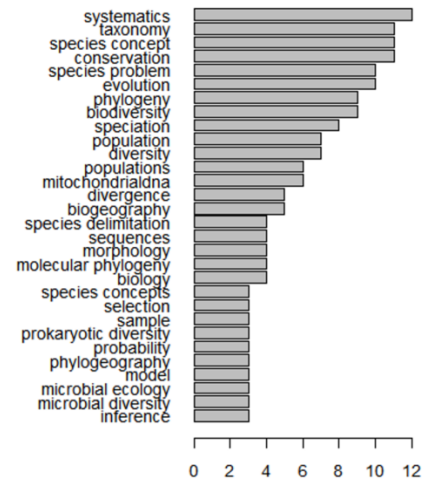




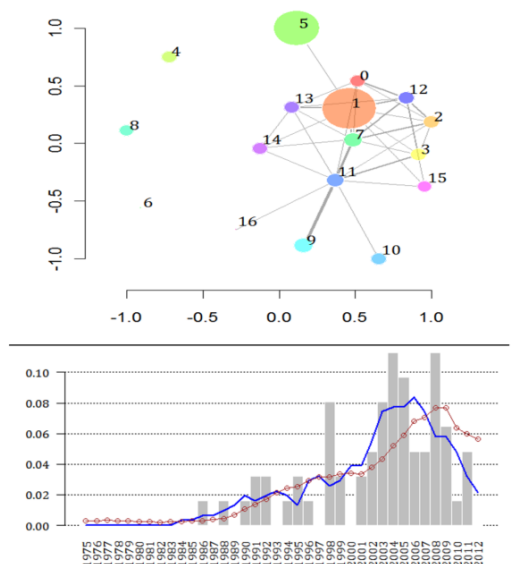


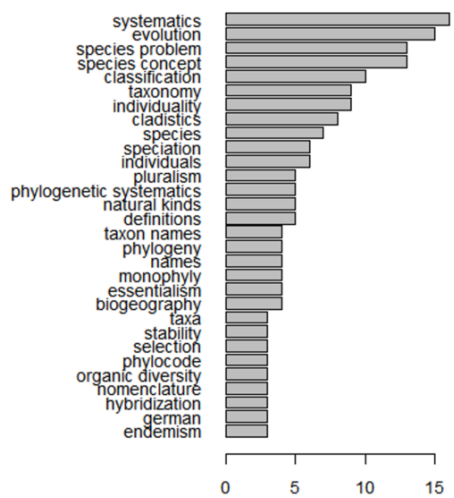
Cluster 7



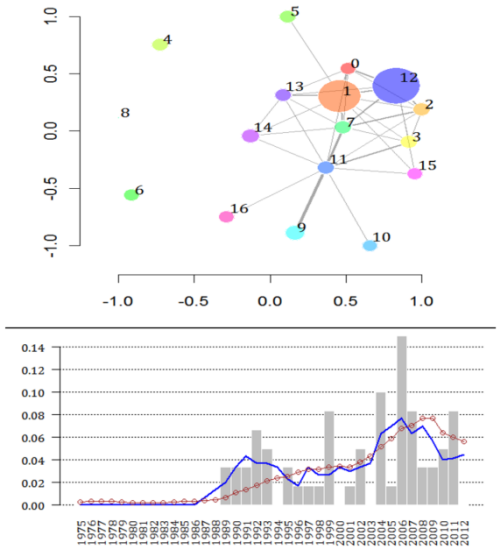


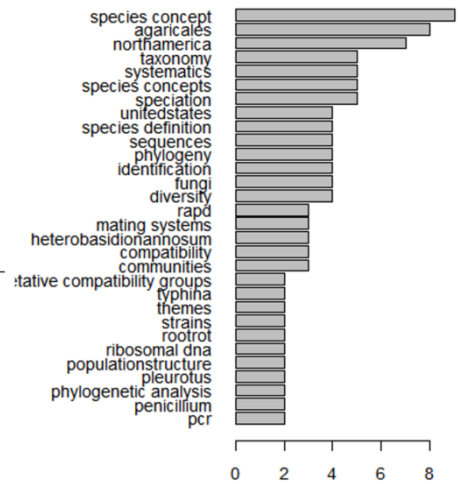
Cluster 8



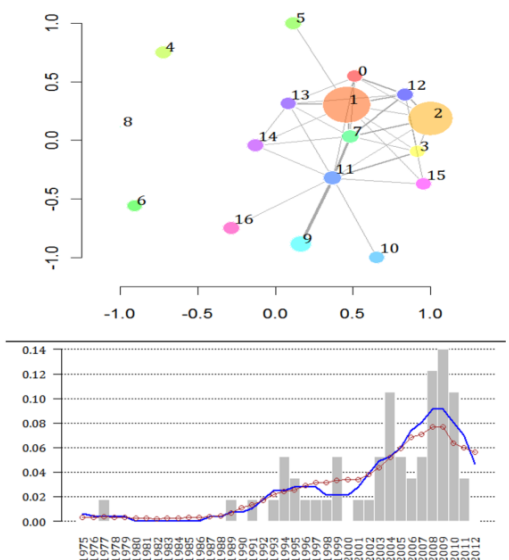


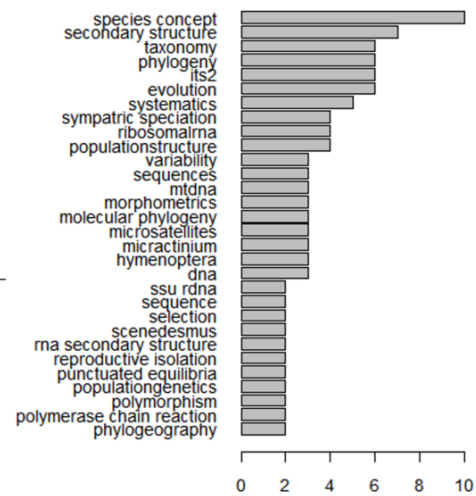
Cluster 9



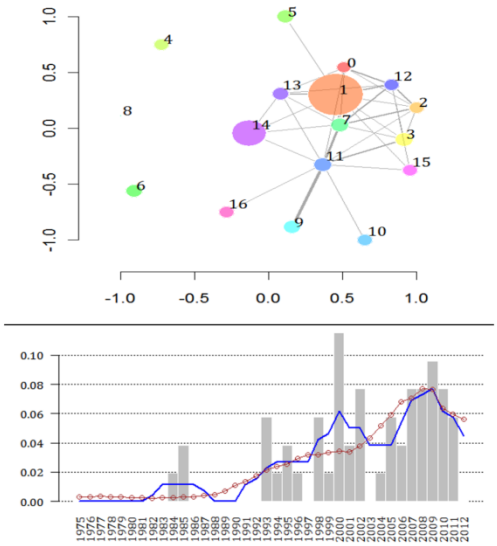


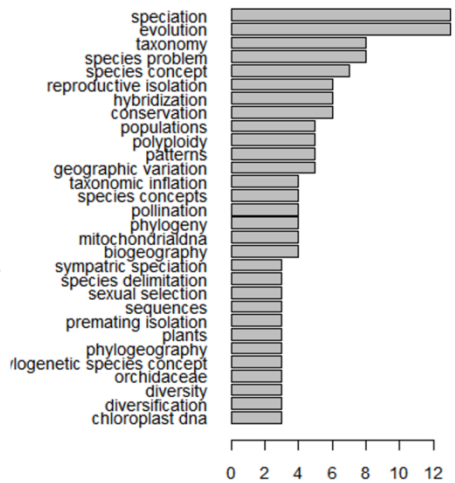
Cluster 10



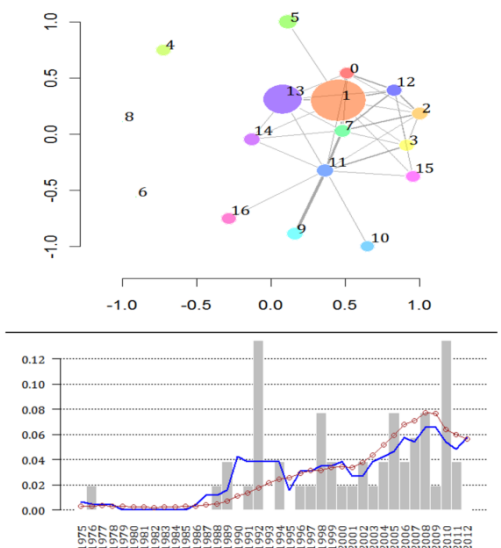


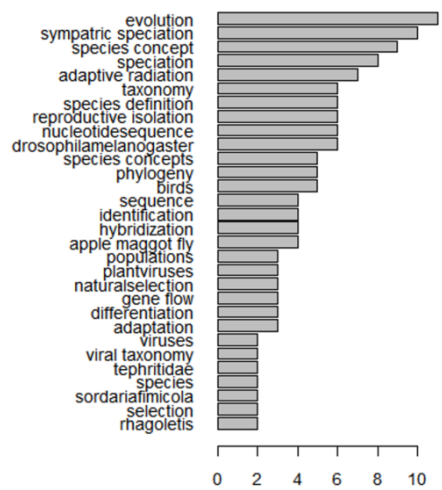
Cluster 11



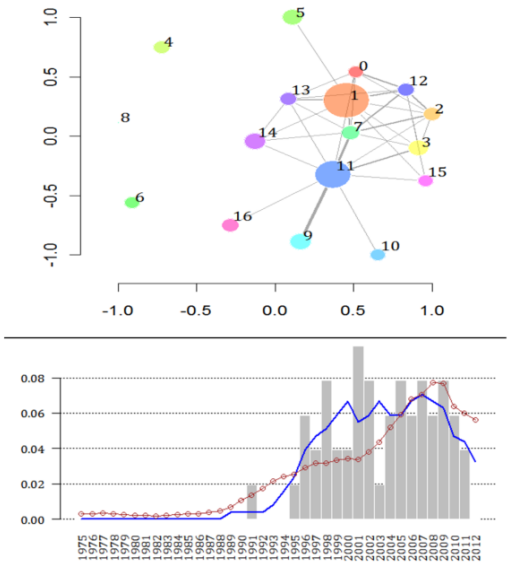


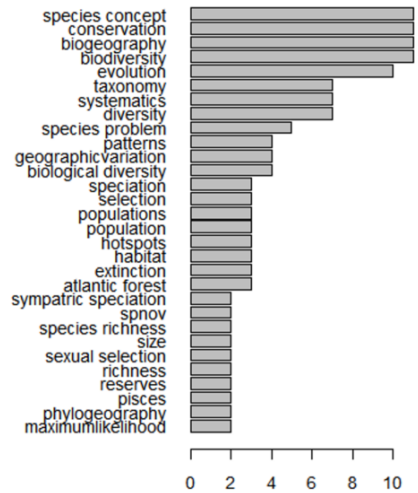
Cluster 12



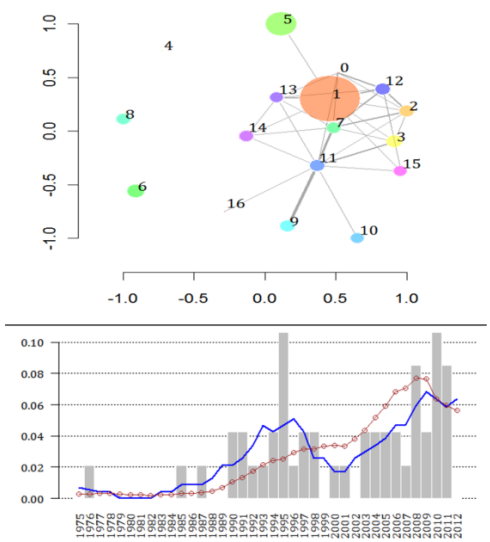


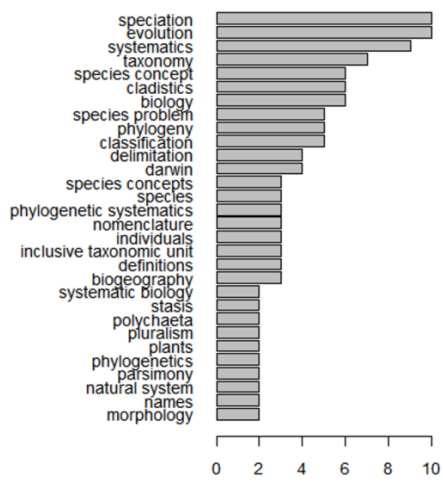
Cluster 13



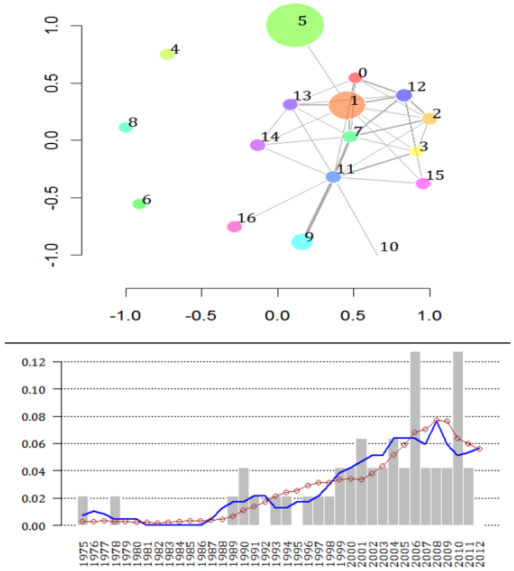


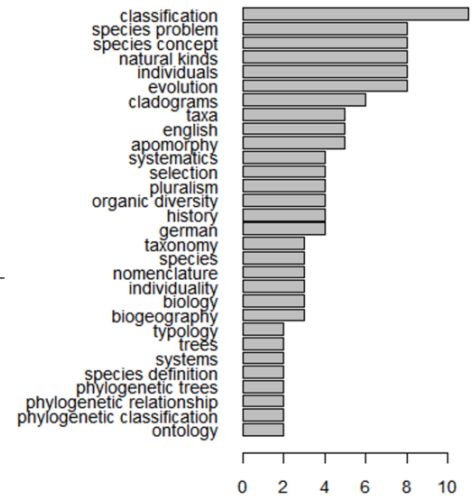
Cluster 14



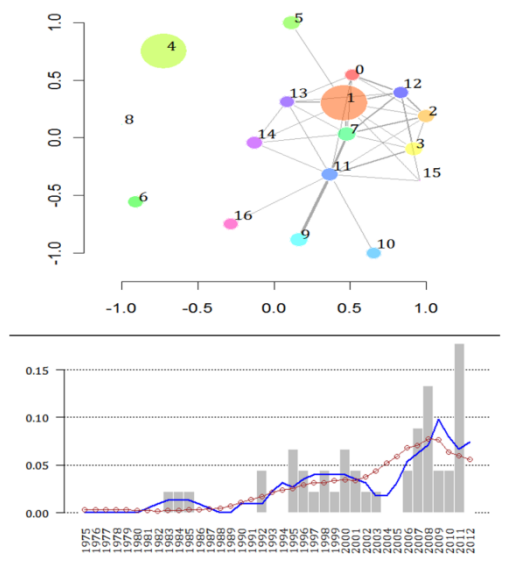


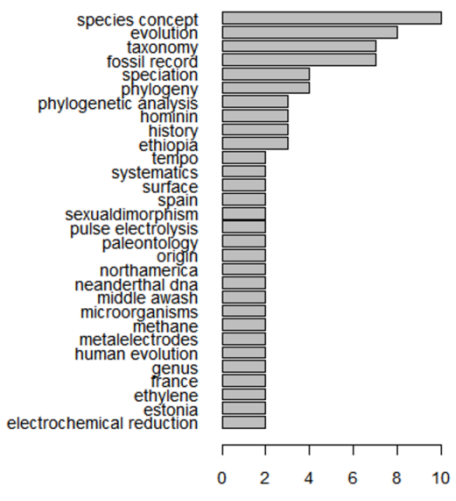
Cluster 15



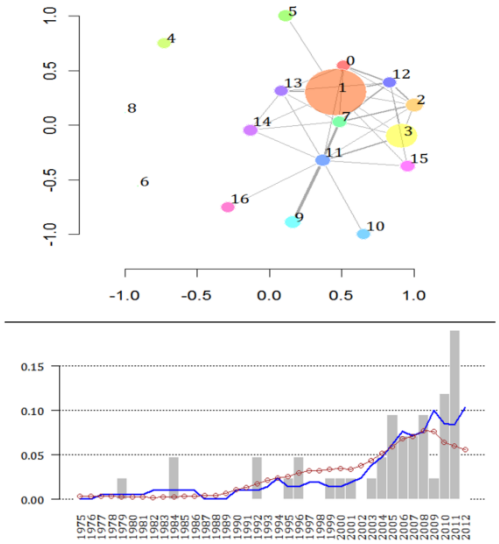


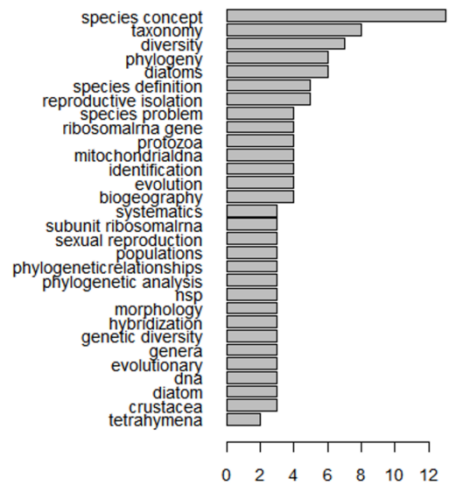
Cluster 16



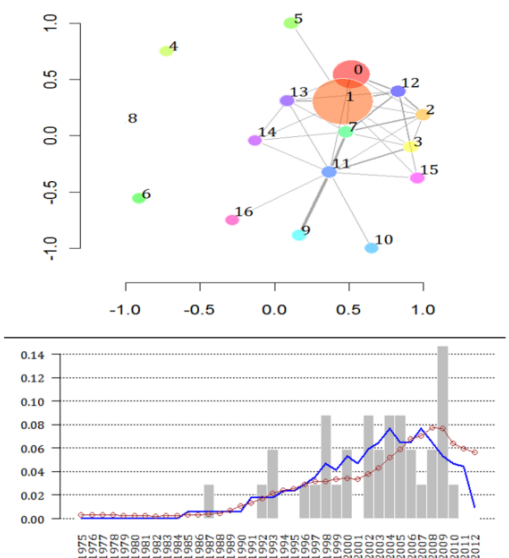


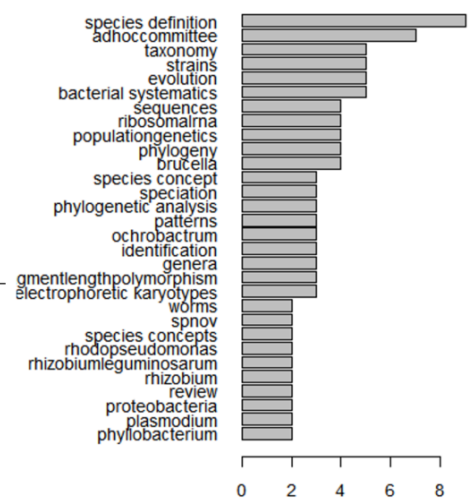
Cluster 17



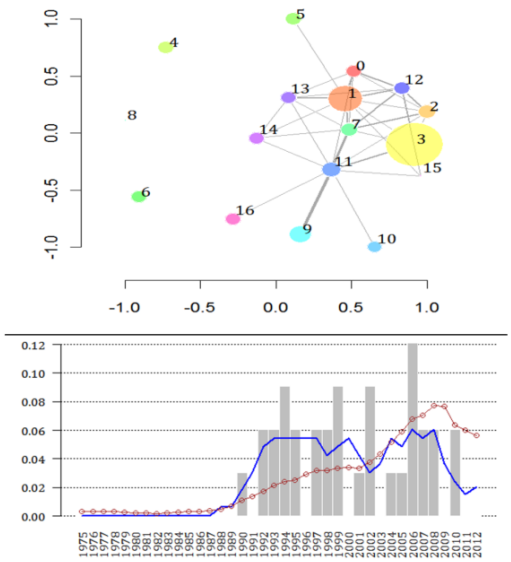


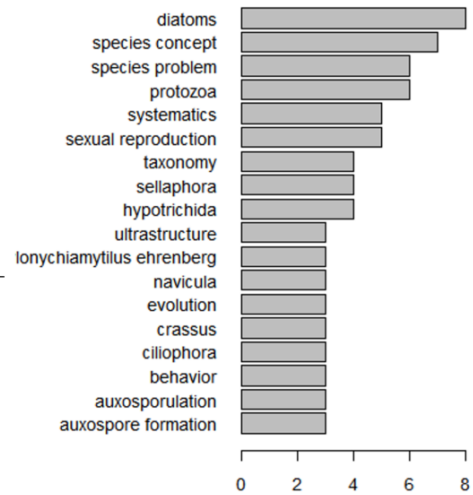
Cluster 18



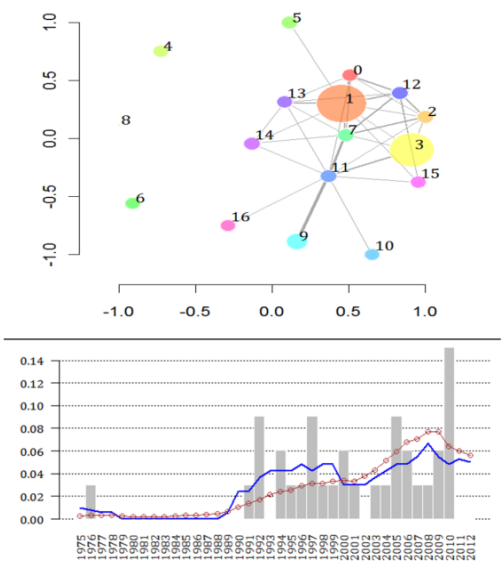


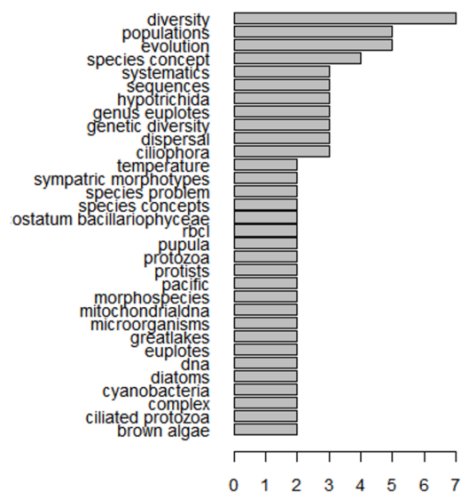
Cluster 19



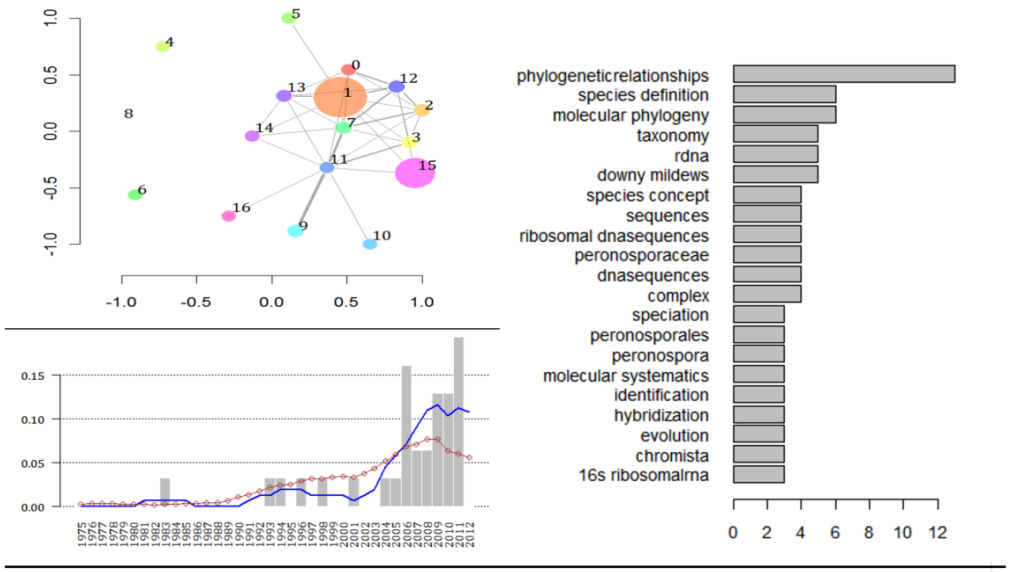


Cluster 20





Cluster 21



Results from model 1.3: Disciplinary interactions

Results of applying the extended overlay toolkit to the citation network of the species problem are presented in Fig 26–28. below. The three plots display the outcome of the three comparative designs discussed above, respectively. Figure 26 shows the time series of *Mean Overlay Distance* values over the time period 1975–2011, measured between (separate) annual source cohorts and their citing environment within the corpus (diversification induced by annual sections of the species topic). Our main focus is Figure 27, whereby the MOD index is calculated via the cumulative method, between the cited side and the citing side being aggregated up to each year within the timespan (diversification *within* the species *topic* by each year). Finally, the longitudinal measurement of the *Overlay Diversity Ratio* is set out in Figure 28, for the purposes of comparison with the MOD measure. The underlying citation relation was also obtained by

the cumulative method, that is, the *Diversity Ratio* was established between cohorts published in year $y \leq Y$ and the citing nodes for y = 1975, …, 2011.

The most striking outcome of this three-way comparison is, in general, that the three curves report a (pairwise) different tendency on the knowledge sharing process, or, more precisely, show a different aspect of the very same process. The cross-sectional perspective (Figure 26) exhibits an initial fluctuation in the MOD index (up to the late eighties), becoming much more moderate, almost steady in later periods, with a sudden peak around 2010. Overall (as a potential smoothing regression of the empirical values would reveal), a descending tendency can be observed in this dimension of citation-based knowledge transfer. This descending tendency is much more heavily present (and differently shaped) in the cumulative perspective (Figure 27). It can be viewed as evidencing, basically, the effect of novel fields entering the topic each year (being absent in previous years) at both the citing and the cited side. MOD values in this application follow a power law-like curve: diversification (through citations) is quickly decreasing up to the early eighties, which tendency continues but slows down in the 80's, and a very low, slightly falling but basically steady level is observable in the last two decades. On the other hand, a radically different picture emerges from measuring the change of diversity between time periods (Figure 28). The ODR index, apart from an initial small oscillation, quickly raises above ODR = 1 and remains within the interval between 1–1.2. In other words, as opposed to the MOD index, knowledge transfer results in an increased diversity of research fields throughout the whole timespan of topic development.

As a brief interpretation of these results, the species problem may be characterized as having a vivid or "revolutionary" period in the 70's–80's opening up interfaces between various and relatively distant fields of research, as evidenced by the MOD index. The cross-sectional analysis suggests that different fields entered the scene in subsequent years, with far-reaching impact relative to each year. The cumulative approach adds, however, that

field composition quickly became "saturated", that is, by accumulating fields along the timeline, the cited and the cited side took an increasingly similar structure. This interpretation is in accord with historiography: according to reconstructions on the history of the topic, a fundamental drive behind the modern debate on the species concept was a thesis from the philosophy of science originating from the mid-seventies and disputed mainly throughout the eighties, that became widely accepted and assimilated within theoretical biology. This so-called *individuality thesis* – stating that species are ontological and methodological individuals – therefore, invited fields such as the *history and philosophy of science* into the discourse otherwise dominated by the life sciences (cited side), and, being a rather influential one, propagated through a variety of fields (w.r.t. the citing side). This extended scope, once being emerged, remained characteristic of the topic in later periods, resulting both (1) little sign of "additional" diffusion and (2) high degree of diversity inherited from this early "boom" of subject areas, as clearly reflected by overlay diversity ratios (ODRs).

The comparison between diversity ratios (Figure 28) and – cumulative – overlay distances (Figure 27) is especially intriguing since both series have been obtained from the same set of cumulative maps. In order to gain a deeper insight into the very process behind the tendencies captured via our proxies, we visualized knowledge transfer at two selected time periods. Respective overlay maps of the cited and the citing side are presented in Figure 29 for the year 1976 and 2001. These two timeslices are quite illustrative as witnessing rather different degrees of diversification (MOD index), but highly similar values of diversity change (ODR index). By consulting the related maps, however, an explanation presents itself.

Source documents published in 1976 were distributed in Subject Categories from mainly the life sciences – Biomedical Sciences, Ecological Sciences, Agricultural Science, Infectious Diseases as disciplines – accompanied with some "non-life" hard sciences (e.g. Geosciences). An area positioned farther

from these fields were “Social Studies”, increasing the distance-based Stirling index for the profile. The associated map indicates that citing papers span a similar field composition, but further Subject Categories, both in the same areas and also in farther regions widen the spectrum of reception: most importantly, a set of fields “mediating” between the “social sciences” and the “natural sciences”, namely, Cognitive Sciences enter the scene with two SC in the middle of the map (increasing the effect of distance in the measurement). In the “social pole”, Business and Management Sciences also pop up. Turning to 2001, both the source map and the target map are much more diversified in this late period, but also much more similar to each other: though novel and relatively distant SCs show up in the citing environment from Computer Science and the collection of “Economics, Politics and Geography”, their share is almost ignorable (below 0,01 percent), so their contribution to the overall share-weighted distance from the source map is almost invisible.

To sum up these effects, knowledge transfer in both years leads to a higher diversity of fields (in terms of the Stirling measure), keeping the ratio of diversities above 1. However, despite of this increment, the overall distance of the citing composition is considerably higher in the early period (described by fewer SCs) than in the later year under study (whereby SCs are abundant). Hence the parallel fall in the MOD measure. This result, beyond explaining the values presented within the time series, provides justification for the use of both approach, as capturing different aspects of the diffusion process.

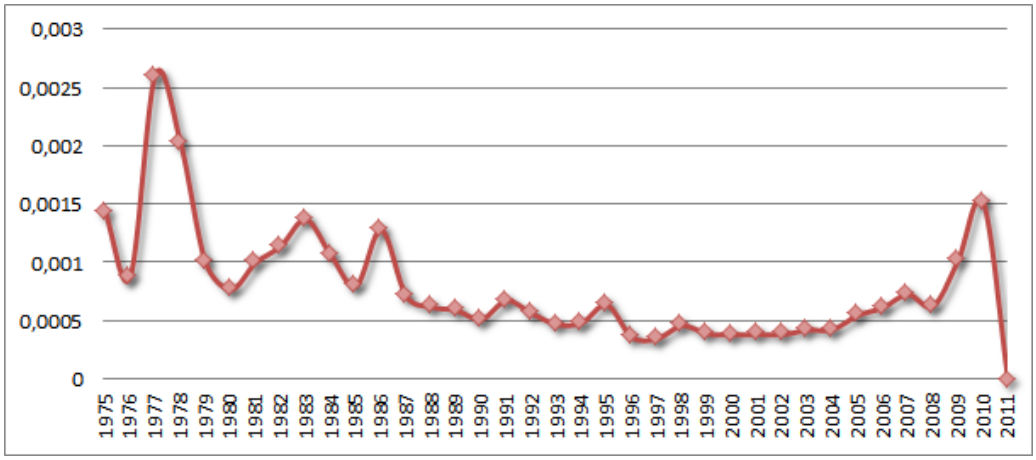


Figure 26. Development of the MOD index comparing annual sections and their citing environment

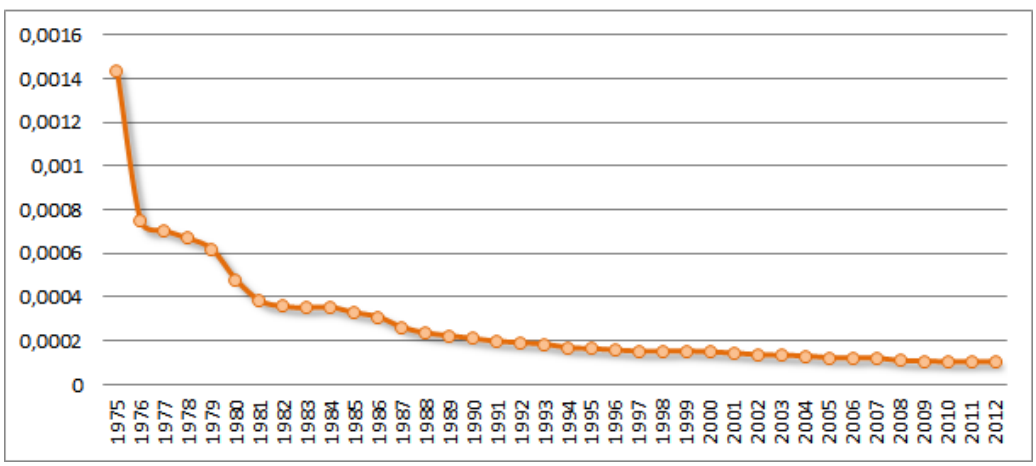


Figure 27. Development of the MOD index comparing accumulated papers wit their citing environment

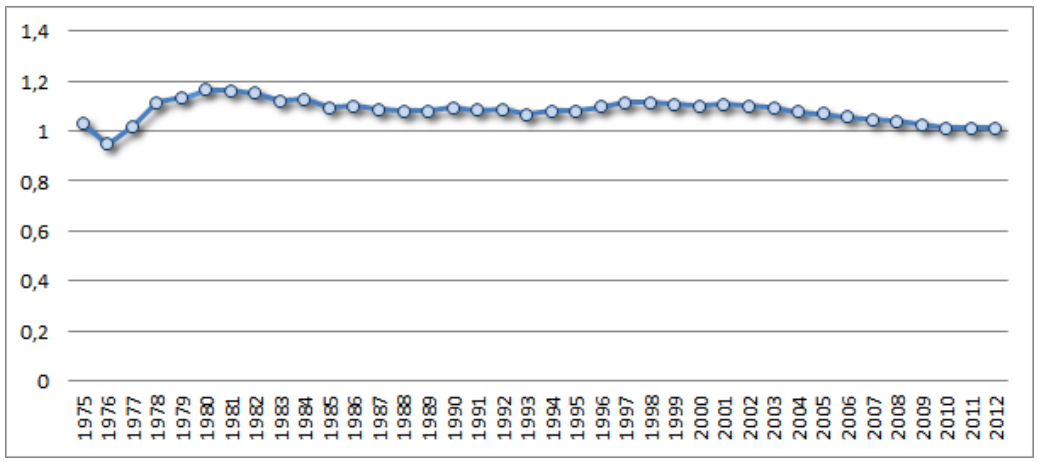


Figure 28. Development of the ODR index comparing the diversity of accumulated papers up to each year with the diversity of their citing environment

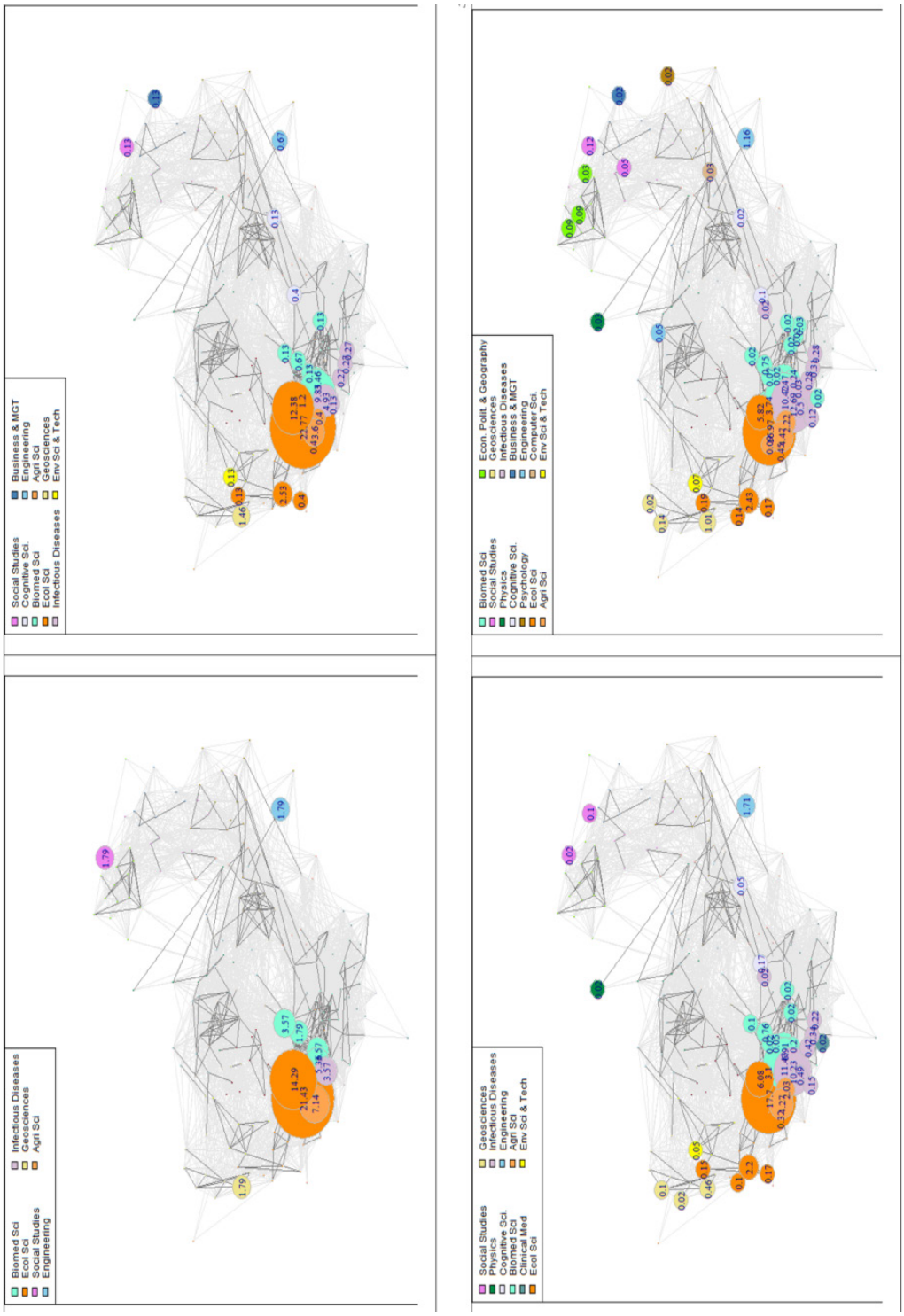


Figure 29. The overlay maps for two selected years visualizing knowledge diffusion between the cited (left) and citing side (right): 1976 (top), 2001 (bottom)

Chapter 8: Conclusion

The widespread view on the species debate holds that the species problem is best interpreted as being present at two levels, in biological science and in the philosophy of science as well. According to such a view, the theoretical–methodological dispute within systematics (that is, biology) is complemented with a parallel dispute in the philosophy of science, reflecting on the philosophical (ontological, conceptual) aspects of the SP. In this perspective, the methodological–scientific line is quite autonomous in the sense that, though clearly interlinked with philosophy at the conceptual level, it is not affected in its development by philosophical arguments.

Despite the intuitive nature of the view on disciplinary autonomy, our analyses of the theoretical debate on the concept of species revealed that the relationship between the two conceptual levels, (1) the scientific-methodological aspect and (2) biophilosophy is not purely conceptual. On the contrary, this relationship plays a causal role in the development of the discourse.

As an empirical confirmation of this causal role, the methods and studies presented in this book provided a positive answer to our third research question (RQ3) by confirming the following pair of hypothesis:

1. In the history of the species problem, biology (biosystematics) have properly incorporated the ontological (philosophical) debate and its implications into its research programme aiming at finding the appropriate species category.

2. As a corollary, throughout the development of the species problem, the distant disciplines participating in and contributing to the debate actually affected each other, that is, there is a causal–historical relationhip between the biological an the philosophical aspects of the debate, partially responsible for the durability of the issue. In other words, the species problem is best

viewed as an interdisciplinary species problem, whereby a deep integration of distant knowledge items shaped the discourse, establishing a case of proper interdisciplinarity.

The mapping of the debate provided evidence that the conceptual problem has been mainly shaped by biological specialities for which mainstream species concepts (such as the Biological Species Concept) implied criteria hardly applicable in the corresponding biological domain. However, from the integration of topics within topical clusters, an equally important drive of the problem can be implied, which is the meta-level (philosophical) debate on the ontological status of both species and the species category. It is confirmed that, though the modern consensus is concerned with molecular methods to discover species, the roadmap to the state-of-the-art accomodates various philosophical "interventions". The roadmap is reflected in the topical cluster profiles: the ontological claim for species taxa being individuals provided philosophical support for the phylogenetic and cladistic concepts of species, which, in turn, served as the theoretical basis for applying methods from molecular genetics to delineate species taxa (reconstructing molecular phylogenies). In fact, all citation-based methods introduced in the previous chapters confirmed the standard story: historical building blocks of the debate involved (1) the polemy on the implications of the Biological Species Concept (2) the attempts to transform its "theoretical potential" phylogenetic and cladistic apporaches by also incorporating ontological claims (Species as Individuals), (3) a philosophical debate expressing itself in strong interactions with systematics, and (4) the role of practical systematics exercised in relation to problematic taxa (microbes, plants, fungi etc.) As such, our analysis confirmed that the history of the species problem is a case of strong interdisciplinarity, whereby distant disciplines interact to form new paradigms.

References

Agapow, P. M., & Sluys, R. (2005). The reality of taxonomic change. *Trends in Ecology & Evolution, 20*(6), 278–280.

Aleman-Meza, B.; Nagarajan, M.; Ramakrishnan, C.; Ding, L.; Kolari, P.; Sheth, A.; Arpinar, I.; Joshi, A. & Finin, T. (2006). Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection, in *'Proceedings of the 15th international conference on World Wide Web'*, pp. 407–416.

Archibald, J. D. (1994). Metataxon concepts and assessing possible ancestry using phylogenetic systematics. S*ystematic Biology, 43*(1), 27–40.

Archibald, J. K., Mort, M. E., & Wolfe, A. D. (2005). Phylogenetic relationships within Zaluzianskya (Scrophulariaceae s.s., tribe Manuleeae): Classification based on DNA sequences from multiple genomes and implications for character evolution and biogeography. *Systematic Botany, 30*(1), 196–215.

Barabási, A.; Jeong, H.; Néda, Z.; Ravasz, E.; Schubert, A. & Vicsek, T. (2002), 'Evolution of the social network of scientific collaborations', *Physica A: Statistical Mechanics and its Applications 311*(3-4), 590–614.

Bettencourt, L.; Kaiser, D. & Kaur, J. (2009), 'Scientific discovery and topological transitions in collaboration networks', *Journal of Informetrics 3*(3), 210–221.

Beurton, P. J. (1995). How is a species kept together. *Biology & Philosophy, 10*(2), 181–196.

Beurton, P. J. (2002). Ernst Mayr through time on the biological species concept – a conceptual analysis. *Theory in Biosciences, 121*(1), 81–98.

Bird, A. (2010). Social knowing: The social sense of'scientific knowledge'. *Philosophical perspectives*, 24, 23–56.

Börner, K. (2010), *Atlas of Science: Visualizing What We Know.* Cambridge, MA: MIT Press.

Boyack, K. (2009), 'Using detailed maps of science to identify potential collaborations', *Scientometrics 79*(1), 27–44.

Boyack, K.; Klavans, R. & Börner, K. (2005), 'Mapping the backbone of science', *Scientometrics 64*(3), 351–374.

Carley, S., & Porter, A. L. (2012). A forward diversity index. *Scientometrics, 90*(2), 407-427.

Ceotto, P. C., & Mejdalani, G. (2005). Phylogenetic analysis of the *Abana* group of genera (Hemiptera : Cicadellidae : Cicadellinae : Proconiini). *Systematic Entomology, 30*(3), 480–496.

Chambers, G. (2012). The species problem: seeking new solutions for philosophers and biologists. *Biology & Philosophy, 27*(5), 755–765.

Chen C. (2013). *Mapping scientific frontiers: the quest for knowledge visualization.* Springer.

Chen, C. (2003), *Mapping Scientific Frontiers: The Quest for Knowledge Visualization.* New York: Springer Verlag.

Chen, C., & Song, M. (2017). *Representing scientific knowledge.* New York: Springer.

Christoffersen, M. L. (1995). Cladistic taxonomy, phylogenetic systematics, and evolutionary ranking. *Systematic Biology, 44*(3), 440–454.

Chu, P. C. (1998). A phylogeny of the gulls (Aves : Larinae) inferred from osteological and integumentary characters. *Cladistics, 14*(1), 1–43.

Chung, C. (2004). The species problem & the value of teaching the complexities of species. *American Biology Teacher, 66*(6), 413–417.

Crane, J. K. (2004). On the metaphysics of species. *Philosophy of Science, 71*(2), 156–173.

Crisp, M. D., & Weston, P. H. (1993). Geographic and ontogenic variation in morphology of Australian Waratahs (Telopea, Proteaceae). *Systematic Biology, 42*(1), 49–76.

Csardi, G. & Nepusz, T. (2006), 'The igraph software package for complex network research', *InterJournal Complex Systems 1695*(1695).

Cusimano, N., Stadler, T., & Renner, S. S. (2012). A New Method for Handling Missing Species in Diversification Analysis Applicable to Randomly or Nonrandomly Sampled Phylogenies. *Systematic Biology, 61*(5), 785–792.

de Moya Anegón, F.; Contreras, E. & Corrochano, M. (1998), 'Research fronts in library and information science in Spain (1985–1994)', *Scientometrics 42*(2), 229–246.

de Queiroz, K. (2005). Different species problems and their resolution. *Bioessays, 27*(12), 1263–1269.

Dequeiroz, K. (1992). Phylogenetic definitions and taxonomic philosophy. *Biology & Philosophy, 7*(3), 295–313.

Dequeiroz, K. (1994). Replacement of an essentialistic perspective on taxonomic definitions as exemplified by the definition of Mammalia. *Systematic Biology, 43*(4), 497–510.

Dequeiroz, K., & Donoghue, M. J. (1988). Phylogenetic systematics and the species problem. *Cladistics-the International Journal of the Willi Hennig Society, 4*(4), 317–338.

Dietrich, M. R. (2004). Genes, categories, and species: The evolutionary and cognitive causes of the species problem. *Philosophy of Science, 71*(4), 619–620.

Dover, G. (1995). A species definition – a functional-approach. *Trends in Ecology & Evolution, 10*(12), 489–490.

Ereshefsky, M. (1992). *The Units of Evolution: Essays on the Nature of Species.* Cambridge: MIT Press.

Ereshefsky, M. (2007). Foundational issues concerning taxa and taxon names. *Systematic Biology, 56*(2), 295–301.

Ereshefsky, M. (2010). Darwin's solution to the species problem. *Synthese, 175*(3), 405–425.

Ereshefsky, M. (2010). Microbiology and the species problem. *Biology & Philosophy, 25*(4), 553–568.

Ereshefsky, M. (2011). Mystery of mysteries: Darwin and the species problem. *Cladistics, 27*(1), 67–79.

Ereshefsky, M., & Matthen, M. (2005). Taxonomy, polymorphism, and history: An introduction to population structure theory. *Philosophy of Science, 72*(1), 1–21.

Garfield, E.; Pudovkin, A. & Istomin, V. (2002), 'Algorithmic citation-linked historiography – Mapping the literature of science', *Proceedings of the American Society for Information Science and Technology 39*(1), 14–24.

Ghiselin, M. T. (1974). Radical solution to species problem. *Systematic Zoology*, 23(4), 536–544.

Giray, E. F. (1976). Integrated biological approach to species problem. *British Journal for the Philosophy of Science, 27*(4), 317–328.

Gittenberger, E. (1995). A species definition – a functional-approach. *Trends in Ecology & Evolution, 10*(12), 490–490.

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. *Scientometrics, 111*, 1071–1087.

Gläser, J. (2020). Opening the black box of expert validation of bibliometric maps. In *Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus*. SoS Discussion Paper, pp. 27–36.

Goldman, A. I. (2011). A guide to social epistemology. In Goldman A. I. & Whitcomb D. (2011). *Social epistemology : essential readings.* Oxford University Press. pp 11–37.

Goyal, S.; van der Leij, M. & Moraga-González, J. (2006), 'Economics: An Emerging Small World', *The Journal of Political Economy 114*(2), 403–412.

Grant, T., & Kluge, A. G. (2004). Transformation series as an ideographic character concept. *Cladistics-the International Journal of the Willi Hennig Society, 20*(1), 23–31.

Griffiths, P. E. (1994). Cladistic classification and functional explanation. *Philosophy of Science, 61*(2), 206–227.

Gutmann, M., & Janich, P. (1998). Species as cultural kinds – Towards a culturalist theory of rational taxonomy. *Theory in Biosciences, 117*(3), 237–288.

Haber, M. H., & Hamilton, A. (2005). Coherence, consistency, and cohesion: Clade selection in Okasha and beyond. *Philosophy of Science, 72*(5), 1026–1040.

Hey, J. (2001). The mind of the species problem. *Trends in Ecology & Evolution, 16*(7), 326–329.

Hey, J., & Pinho, C. (2012). Population genetics and objectivity in species diagnosis. *Evolution, 66*(5), 1413–1429.

Hey, J., Waples, R. S., Arnold, M. L., Butlin, R. K., & Harrison, R. G. (2003). Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology & Evolution, 18*(11), 597–603.

Hjørland, B. (2013). Citation analysis: A social and dynamic approach to knowledge organization. *Information Processing & Management, 49*(6), 1313–1325.

Hull D. L. (1990). *Science as a process: an evolutionary account of the social and conceptual development of science (science and its conceptual foundations)*. University of Chicago Press.

Hull, D. L. (2001). Michael Ruse and his fifteen years of Booknotes – For better or for worse. *Biology & Philosophy, 16*(3), 423–435.

Hull, D.L. (1978). Matter of individuality. *Philosophy of Science*, *45*(3), 335–360.

Hull, D.L. (1988). *Science as a process: an evolutionary account of the social and conceptual development of science.* Chicago: University of Chicago Press.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10–25.

Klavans, R. & Boyack, K. (2007), Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases, in *'Proceedings of ISSI'*, pp. 437–48.

Kouwets, F. A. C. (2008). The species concept in desmids: the problem of variability, infraspecific taxa and the monothetic species definition. *Biologia, 63*(6), 881–887.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719–720.

LaPorte, J. (1997). Essential membership. *Philosophy of Science, 64*(1), 96–112.

LaPorte, J. (2003). Genes, categories, and species: The evolutionary and cognitive causes of the species problem. *British Journal for the Philosophy of Science, 54*(4), 627–630.

LaPorte, J. (2006). The species problem: Biological species, ontology, and the metaphysics of biology. *Biology & Philosophy, 21*(3), 381–393.

Levine, A. (2001). Individualism, type specimens, and the scrutability of species membership. *Biology & Philosophy, 16*(3), 325–338.

Lewens, T. (2012). Pheneticism reconsidered. *Biology & Philosophy, 27*(2), 159–177.

Leydesdorff, L. & Rafols, I. (2009), 'A global map of science based on the ISI subject categories', *Journal of the American Society for Information Science and Technology 60*(2), 348–362.

Leydesdorff, L. (1997), 'Why words and co-words cannot map the development of the sciences', *Journal of the American Society for Information Science 48*(5), 418–427.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87–100.

Liden, M. (1990). Replicators, hierarchy, and the species problem. *Cladistics-the International Journal of the Willi Hennig Society, 6*(2), 183–186.

Lockwood, J. A. (2012). Species are Processes: A Solution to the 'Species Problem' via an Extension of Ulanowicz's Ecological Metaphysics. *Axiomathes, 22*(2), 231–260.

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive psychology, 32*(3), 251–295.

Los, B. (2000). The empirical performance of a new inter-industry technology spillover measure. *Technology and Knowledge*, 118–151.

Lovtrup, S. (1987). On the species problem and some other taxonomic issues. *Environmental Biology of Fishes, 20*(1), 3–9.

Mallet, J. (1995). A species definition – a functional-approach – reply. *Trends in Ecology & Evolution, 10*(12), 490–491.

Mallet, J. (1995). A species definition for the modern synthesis. *Trends in Ecology & Evolution, 10*(7), 294–299.

Mayden, R. L. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem. In M. F. Claridge, H. A. Dawah & M. R. Wilson (Eds.), *Species, the units of biodiversity* (pp. 381–424): Chapman and Hall, London.

Mayr, E. (1996). What is a species, and what is not? *Philosophy of Science, 63*(2), 262–277.

McOuat, G. (2001). Cataloguing power: delineating 'competent naturalists' and the meaning of species in the British Museum. *British Journal for the History of Science, 34*(120), 1–28.

McOuat, G. (2001). From cutting nature at its joints to measuring it: New kinds and new kinds of people in biology. *Studies in History and Philosophy of Science, 32A*(4), 613–645.

McOuat, G. R. (1996). Species, rules and meaning: The politics of language and the ends of definitions in 19th century natural history. *Studies In History And Philosophy of Science, 27*(4), 473–519.

Moore, J. A. (1991). Science as a way of knowing. 7. A conceptual-framework for biology. 2. *American Zoologist, 31*(2), 349–470.

Moya-Anegón, F.; Vargas-Quesada, B.; Herrero-Solana, V.; Chinchilla-Rodriguez, Z.; Corera-Álvarez, E. & Munoz-Fernández, F. (2004), 'A new technique for building maps of large scientific domains based on the cocitation of classes and categories', *Scientometrics 61*(1), 129–145.

Muhlenbach, F. & Lallich, S. (2010),Discovering Research Communities by Clustering Bibliographical Data, in *'Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on'*, pp. 500–507.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577.

Ohara, R. J. (1993). Systematic generalization, historical fate, and the species problem. *Systematic Biology, 42*(3), 231–246.

Ohara, R. J. (1994). Evolutionary history and the species problem. *American Zoologist, 34*(1), 12–22.

Oskolski, A. (2011). The Taxon as an Ontological Problem. *Biosemiotics, 4*(2), 201–222.

Pedroso, M. (2013). The Species Problem: A Philosophical Analysis. *Mind, 122*(488), 1180–1182.

Pigliucci, M. (2003). Species as family resemblance concepts: the (dis-) solution of the species problem? *Bioessays, 25*(6), 596–602.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. *Computer and Information Sciences-ISCIS 2005*, 284–293.

Porter, A. L., Cohen, A. S., Roessner, J. D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics, 72*(1), 117–147.

Price, D. J. (1970). Citation measures of hard science, soft science, technology, and nonscience. In C. E. Nelson & D. K. Pollack (Eds.), *Communication among scientists and engineers* (pp. 3–22): Heath, Lexington, MA, USA.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871–1887.

Reed, E. S. (1979). Role of symmetry in Ghiselin radical solution to the species problem. *Systematic Zoology, 28*(1), 71–78.

Reeves, P. A., & Richards, C. M. (2007). Distinguishing terminal monophyletic groups from reticulate taxa: Performance of phenetic, tree-based, and network procedures. *Systematic Biology, 56*(2), 302–320.

Reydon, T. A. C. (2004). Why does the species problem still persist? *Bioessays, 26*(3), 300–305.

Reydon, T. A. C. (2008). Species in three and four dimensions. *Synthese, 164*(2), 161–184.

Reydon, T. A. C. (2009). Species and kinds: a critique of Rieppel's "one of a kind" account of species. *Cladistics, 25*(6), 660–667.

Reydon, T. A. C. (2013). Classifying Life, Reconstructing History and Teaching Diversity: Philosophical Issues in the Teaching of Biological Systematics and Biodiversity. *Science & Education, 22*(2), 189–220.

Reydon, T. A. C. (2013). The Species Problem: A Philosophical Analysis. *Science & Education, 22*(2), 381–389.

Ridley, M. (1989). The cladistic solution to the species problem. *Biology & Philosophy, 4*(1), 1–16.

Ridley, M. (1990). Ridley cladistic solution to the species problem – comment. *Biology & Philosophy, 5*(4), 447–450.

Rieppel, O. (1991). Things, taxa and relationships. *Cladistics-the International Journal of the Willi Hennig Society, 7*(1), 93–100.

Rojas, M. (1992). The species problem and conservation – what are we protecting. *Conservation Biology, 6*(2), 170–178.

Rousseau, R. (1994), 'Similarities between informetrics and econometrics', *Scientometrics 30*(2), 385–387.

Ruse, M. (1995). *The species problem.* Pittsburgh: Univ Pittsburgh Press.

Schilthuizen, M. (2000). Dualism and conflicts in understanding speciation. *Bioessays, 22*(12), 1134–1141.

Scopece, G., Cozzolino, S., & Bateman, R. M. (2010). Just what is a genus? Comparing levels of postzygotic isolation to test alternative taxonomic hypotheses in Orchidaceae subtribe Orchidinae. *Taxon, 59*(6), 1754–1764.

Scopece, G., Musacchio, A., Widmer, A., & Cozzolino, S. (2007). Patterns of reproductive isolation in Mediterranean deceptive orchids. *Evolution, 61*(11), 2623–2642.

Sicard, M., Desmarais, E., & Lambert, A. (2001). Molecular characterisation of Diplozoidae populations on five Cyprinidae species: consequences for host specificity. *Comptes Rendus De L Academie Des Sciences Serie Iii-Sciences De La Vie-Life Sciences, 324*(8), 709–717.

Skinner, A. (2004). Hierarchy and monophyly. *Cladistics-the International Journal of the Willi Hennig Society, 20*(5), 498–500.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265–269.

Smith, J. M., Feil, E. J., & Smith, N. H. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays, 22*(12), 1115–1122.

Soós, S. (2008). *A tudományos és a naiv fogalomrendszer kölcsönhatásának vizsgálata a fajproblematika [species problem] interdiszciplináris modelljének rekonstrukcióján keresztül.* PhD dissertation, ELTE PPK.

Soós, S., & Kampis, G. (2011). Towards a typology of research performance diversity: the case of top Hungarian players. *Scientometrics, 87*(2), 357–371.

Soós, S., & Kampis, G. (2012). Beyond the basemap of science: mapping multiple structures in research portfolios: evidence from Hungary. *Scientometrics, 93*(3), 869–891.

Stamos, D. N. (1996). Popper, falsifiability, and evolutionary biology. *Biology & Philosophy, 11*(2), 161–191.

Stamos, D. N. (2002). Species, languages, and the horizontal/vertical distinction. *Biology & Philosophy, 17*(2), 171–198.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719.

Takacs, P., & Ruse, M. (2013). The Current Status of the Philosophy of Biology. *Science & Education, 22*(1), 5–48.

Thagard, Paul (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*, Cambridge, MA: MIT Press.

van Raan, A. F. J. (2005). Reference-based publication networks with episodic memories. *Scientometrics, 63*(3), 549–566.

Vanderpoorten, A., & Goffinet, B. (2006). Mapping uncertainty and phylogenetic uncertainty in ancestral character state reconstruction: An example in the moss genus Brachytheciastrum. *Systematic Biology, 55*(6), 957–971.

Vrana, P., & Wheeler, W. (1992). Individual organisms as terminal entities – laying the species problem to rest. *Cladistics-the International Journal of the Willi Hennig Society, 8*(1), 67–72.

Wang D. & Barabási Albert-László. (2021). *The science of science.* Cambridge University Press.

Wang, W., Do, D. B., & Lin, X. (2005). Term graph model for text classification. In *Advanced Data Mining and Applications* (pp. 19–30): Springer.

Wheeler, Q. D., & Nixon, K. C. (1990). Another way of *looking at* the species problem – a reply to de Queiroz and Donoghue. *Cladistics-the International Journal of the Willi Hennig Society, 6*(1), 77–81.

Wilkinson, M. (1990). A commentary on Ridley cladistic solution to the species problem. *Biology & Philosophy, 5*(4), 433–446.

Wilson, B. E. (1995). A (not-so-radical) solution to the species problem. *Biology & Philosophy, 10*(3), 339–356.

Wilson, B. E. (1995). *"The species problem" – Comments.* Pittsburgh: Univ Pittsburgh Press.

Zabell, S. L. (1992). Predicting the unpredictable. *Synthese, 90*(2), 205–232.

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics, 93*(3), 787–812.