Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge

Quentin Lobbé, Alexandre Delanoë, David Chavalarias

Journal Title
XX(X):1–18

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge

Quentin Lobbé, Alexandre Delanoë and David Chavalarias

Abstract

The ICT revolution has given birth to a world of digital traces. A wide number of knowledge-driven domains like science are daily fueled by unlimited flows of textual contents. In order to navigate across these growing constellations of words, interdisciplinary innovations are emerging at the crossroad between social and computational sciences. In particular, complex systems approaches make it now possible to reconstruct multi-level and multi-scale dynamics of knowledge by means of inheritance networks of elements of knowledge called *phylomemies*.

In this article, we will introduce an endogenous way to visualize the multi-level and multi-scale properties of phylomemies. The resulting system will enrich a state-of-the-art tree like representation with the possibility to browse through the evolution of a corpus of documents at different level of observation, to interact with various scales of description, to reconstruct a hierarchical clustering of elements of knowledge and to navigate across complex semantic lineages. We will then formalize a generic macro-to-micro methodology of exploration and implement our system as a free software called the *Memiescape*. Our system will be illustrated by three use cases that will respectively reconstruct the scientific landscape of the top cited publications of the french CNRS, the evolution of the state of the art of *knowledge dynamics visualization* and the ongoing discovery process of Covid-19 vaccines.

Keywords

phylomemy reconstruction, knowledge dynamics, multi-level, multi-scale, science map, co-word analysis

1 Introduction

Since the dawn of humanity, writing has been one of the first mnemotechnology: a technique not only designed to fix a thought on a medium but also a dynamic tool for the elaboration of a collective memory[1]. Written texts can thus be considered as vectors of knowledge as well as providers of socio-historical contexts. The accumulation of Mesopotamian clay tablets (4000 BC) or the elaboration of the Vivarium library (535 AD-555 AD) gave early evidence of a growing will to collect and provide access to isolated elements of knowledge. Later on, with the transition from manuscript to book[2], textual contents outgrew erudite communities and started to touch all layers of the population, up to the present day: we are now daily fueled by unlimited flows of articles, novels, messages, tweets, etc. The recent ICT revolution[3] has given birth to an unprecedented world of digital traces and has impacted a wide number of knowledge-driven domains such as education or policy making.

Science, in particular, has been one of the first area to experiment this digital shift. Databases of scientific publications are scaling up and it is now possible to dive into the amazing richness of most of these catalogs. Qualitative sciences are also taking advantage of the ICT revolution by integrating large cultural data sets (digitized historical documents, social networks footprints, archived Web sites, etc.) within their own scopes of analysis[4]. Digital-born fields of research have thus emerged at the crossroad between social and computational sciences. But whether we speak of digital humanities[5] or cultural analytics[6], it seems that all these domains end up facing the same issue: how to navigate across growing constellations of words and texts?

As early as the 18th century, while he was completing the first edition of the *Encyclopédie*[7], d'Alembert suggested the idea of using trees to situate the future reader:

> *"[...] above this vast labyrinth, whence he can perceive the principal sciences and the arts simultaneously. From there he can see at a glance the objects of their speculations and the operations which can be made on these objects; he can discern the general branches of human knowledge, the points that separate or unite them; and sometimes he can even glimpse the secrets that relate them to one another [...]"*[*]

Likewise, at the beginning of the 1900s, Paul Otlet planned to use experimental interfaces for the legendary *Mundaneum*[8]. The visitors of the library were supposed to have access to the collected documents' references by means of macro-visualizations[†] placed on top of mobile pieces of

[*] d'Alembert, Jean-Baptiste le Rond. "Preliminary Discourse." The Encyclopedia of Diderot & d'Alembert Collaborative Translation Project. Translated by Richard N. Schwab and . Ann Arbor: Michigan Publishing, University of Michigan Library, 2009. Web. 2021. http://hdl.handle.net/2027/spo.did2222.0001.083. Trans. of "Discours Préliminaire," Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, vol. 1. Paris, 1751.

[†] https://upload.wikimedia.org/wikipedia/commons/2/29/Mondoth%C3%A8que_02.jpg

furniture. These few examples are evidence of a historical need: researchers have always been looking for tools able to visualize the dynamics and structures of wide elements of knowledge. In line with D'Alembert's vision, our article's purpose is to give researchers the means to interact with large corpora of texts in relevance with their investigations. To that end, we will rely here on the most recent developments in the fields of *collective intelligence* and *reconstruction methods*.

Collective shapes of knowledge. Because there are structures inside knowledge, a given text can always be studied in relation to others or in light of a specific socio-cultural context. By way of textual traces, human beings are calling out to one another: citations, retweets, controversy, etc. We are the architects of a giant web of elements of knowledge whose very structures and shapes convey information of their own[9]. Like ants or bees, through the aggregation of individual contributions, we collectively achieve complex artifacts that are out of the reach of individuals. This phenomenon is called *collective intelligence*[10] and relies on a core mechanism called *stigmergy*; that is, the indirect coordination between an agent and an action through the environment[11]. From scientific archives to Web pages and online ratings, our digital societies are literally embedded in a stigmergic environment. Wikipedia, for instance, is an emblematic example[12]. The global shapes of these traces of collective intelligence constitute a full-fledged source of knowledge.

Reconstruction methods. Nowadays, complex systems approaches enable us to reconstruct the collective shapes and ontogeny of large corpora of texts. We call *reconstruction methods* all techniques implemented to understand a complex object or natural phenomenon by means of both the observation of patterns and the analysis of processes. Such methods are part of the larger family of *phenomenological reconstructions*, designed to find reasonable approximations of the structure and dynamics of a given phenomenon[13]. Reconstruction methods can be summarized by the generic workflow $O \in \mathcal{O} \rightarrow R \in \mathcal{R} \rightarrow V \in \mathcal{V}$, where O represents a complex phenomenon associated to a set of properties. Based on a collected data set, O is next reconstructed as a formal object R described in a high-dimensional space on the basis of a collected data set. The process ends with the dimensional reduction of R, so that it can be projected as a human-readable visualization V. In this paper, we focus on the *phylomemy reconstruction* method (see B) which consists in reconstructing inheritance networks of elements of knowledge on top of timestamped corpora of textual documents[14]. By means of phylomemies we are now able to reconstruct the dynamics of knowledge, but the question of their visualization remains an open challenge.

Multi-level and multi-scale. Recent research have investigated the multi-level and multi-scale properties of phylomemies (in $\mathcal{R}$)[15] and have established a clear distinction between these two notions. By choosing a specific *level*, we determine the range of intrinsic complexity of the dynamic entity we want to observe. By choosing a specific *scale*, we define the complexity of the description of this entity. While the *scale of description* does not necessarily imply time, the *level of observation* ontologically influences the temporal relations and the structure of the targeted phenomenon since all its elements derive their unity from some underlying dynamic processes.

In order to project a phylomemy from $\mathcal{R}$ to $\mathcal{V}$, our paper will introduce an endogenous system of visualization which relies on their multi-level and multi-scale properties. As the articulation between *levels* and *scales* has not been yet investigated by the *information visualization* community (see 2.1), we thus think our work can contribute to the whole domain of *knowledge dynamics visualization*.

Our main challenge will be to extract the semantic and temporal information embodied within the complex structure of a phylomemy and translate it in a graphical way. As an illustration, the Figure 1 represents, in a discrete space $\mathcal{R}$, the whole structure of a phylomemy on *knowledge mapping* literature. This phylomemy features a set of reconstructed fields of knowledge (yellow dots) and all their possible kinship links (blue to white lines) parameterized by a tolerance thresholds λ for inter-temporal matching. This structure is non-human-readable as seen: its complexity restricts its understanding because there are too many ways to define the parents / children of a given field. The phylomemy thus needs to be simplified by the choice of a level of observation that will decide, for each field, what is the appropriate definition of parents / children.

In what follows, we will describe a tree like visualization designed to represent the basic properties of phylomemies at a given level of observation: inheritance lineages, terms dynamics, branches evolution, etc. How can we combine micro-to-macro graphical elements and interactions mechanisms in a single view? We will enrich this view with multi-level features able to give insights into the meta-structure surrounding it at lower / higher levels. Can we foreshadow a more precise temporal organization? Will this branch split into sub-domains if we vary the level of observation? How can we visualize the hierarchical branches' layout induced by the choice of a given level? In addition, we will investigate a way to combine this multi-level information with multi-scale mechanisms. Indeed, various sets of synchronic scales of description are hidden within the Figure 1. Is there an endogenous way to simplify the representation of a targeted branch of knowledge without additional calculation? Can we fold / unfold on-the-fly its semantic content and temporal complexity?

Summary. In section 2, we will first review the state of the art of *text analysis* and *knowledge visualization*. We will position our approach before comparing it to a selection of related papers (see Table 1). We will end this section by describing the *sea level rise* algorithm (see 2.2.1): one of the core mechanisms of the phylomemy reconstruction process which has influenced the design of our visualization system. With this in mind, we will review in section 3 the key features of our contribution along with their technical implementations. We will thus explain the way we succeed in graphically translating the multi-level and multi-scale properties of phylomemies. In section 4, we will put our system into practice by means of three different use cases: with 4.1 we will give a standard walk-through of the exploration of a corpus of scientific publications; with 4.2 we will validate our approach by browsing through the evolution of the state of the art formerly introduced in section 2; with 4.3 we will report a concrete user experiment. In section 5,

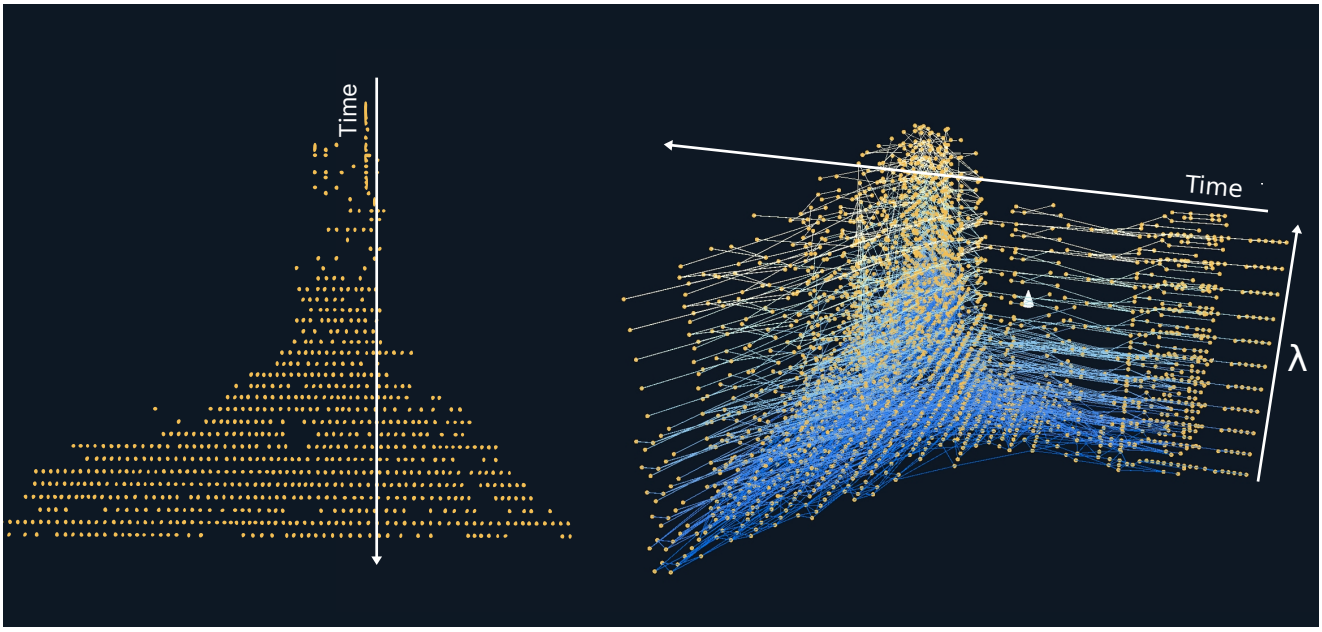

Figure 1. Example of a phylomemy represented in a discrete space $\mathcal{R}$ (reconstructed from the corpus $\mathcal{D}_{maps}$). Each dot on the left represents a cluster of terms describing a field of research at a given period. The phylomemy (on the right) features all these fields and all possible kinship links (blue to white lines) between them parameterized by a tolerance thresholds λ for inter-temporal matching. This threshold, that belongs to the continuous interval $[0\ 1]$, has been discretized at the values $\{0, 0.1, 0.2, ..., 1\}$. Mathematically speaking, the phylomemy on the right is a foliation over the temporal clustering represented on the left. Interactive version available at
http://maps.gargantext.org/unpublished_maps_phylo/phylo_knowledge_3D/index.html

we will discuss the ways to integrate our visualization system into a more generic methodology for revealing the multi-level and multi-scale dynamics of knowledge. We will finally end this paper in section 6 by foreshadowing the benefits of integrating to our system the reflexive and collaborative features of the free software *Gargantext*.

Experimental materials. In section 2, 3 and 4, we will use three different corpora downloadable in C to illustrate the visualization of various phylomemies:

1. The corpus $\mathcal{D}_{cnrs}$ is a collection of 6000 top cited papers extracted from the *Web of Science* (WoS). Written between the '80s and the present days, at least one of each publication's authors is affiliated to the french CNRS. By construction, $\mathcal{D}_{cnrs}$ is an heterogeneous corpus as it gathers a wide variety of research fields.

2. The corpus $\mathcal{D}_{maps}$ is a collection of 13844 scientific publications related to the field of *text analysis and knowledge visualization*, extracted from the *Web of Science* (WoS) and published between the '80s and the present days. By construction, $\mathcal{D}_{maps}$ is thematically coherent as it focuses on a specific domain.

3. The corpus $\mathcal{D}_{ct}$ is a collection of 1263 timestamped descriptions of clinical trials related to the Covid-19 vaccines and published between March 2020 and June 2021. This corpus gives us the possibility to visualize phylomemies reconstructed on the basis of unstructured texts and finer time frequency (week by week instead of year by year with $\mathcal{D}_{cnrs}$ and $\mathcal{D}_{maps}$).

2 State of the art

Mapping the dynamics of large corpora of texts is an interdisciplinary domain of research that has expended under the influence of the data revolution [16]. In order to position our work, we will first adopt a *co-word analysis* approach [17–19] to map out the overall scientific literature of *text analysis and knowledge visualization*. We will then review in depth a collection of related publications and thus highlight the way this paper contributes to the visualization of knowledge dynamics. To that end, let us start by framing the domain with a complex query based on generic terms such as *science map*, *information cartography*, *knowledge map*, etc. By using this query (see details in A), we shape the corpus $\mathcal{D}_{maps}$ out of 13844 documents meta-data (titles and abstracts) extracted from the *Web of Science* (Wos) and create the semantic map: Figure 2.

2.1 The scientific landscape of text analysis and knowledge visualization

In the resulting Map 2, five main communities of research (the numbered dense areas of terms) appear. We here make the assumption that a scientific community can be represented by the use of specific and shared elements of vocabulary. We review these communities by considering their most representative publications:

1. **Citation analysis.** It was the core of *scientometrics* in the 1970s [21] and was used as a method to assess *scientific impact*. Although focusing on *knowledge domains*, it never really dealt with vizualisation and left it to the *bibliometrics* community.

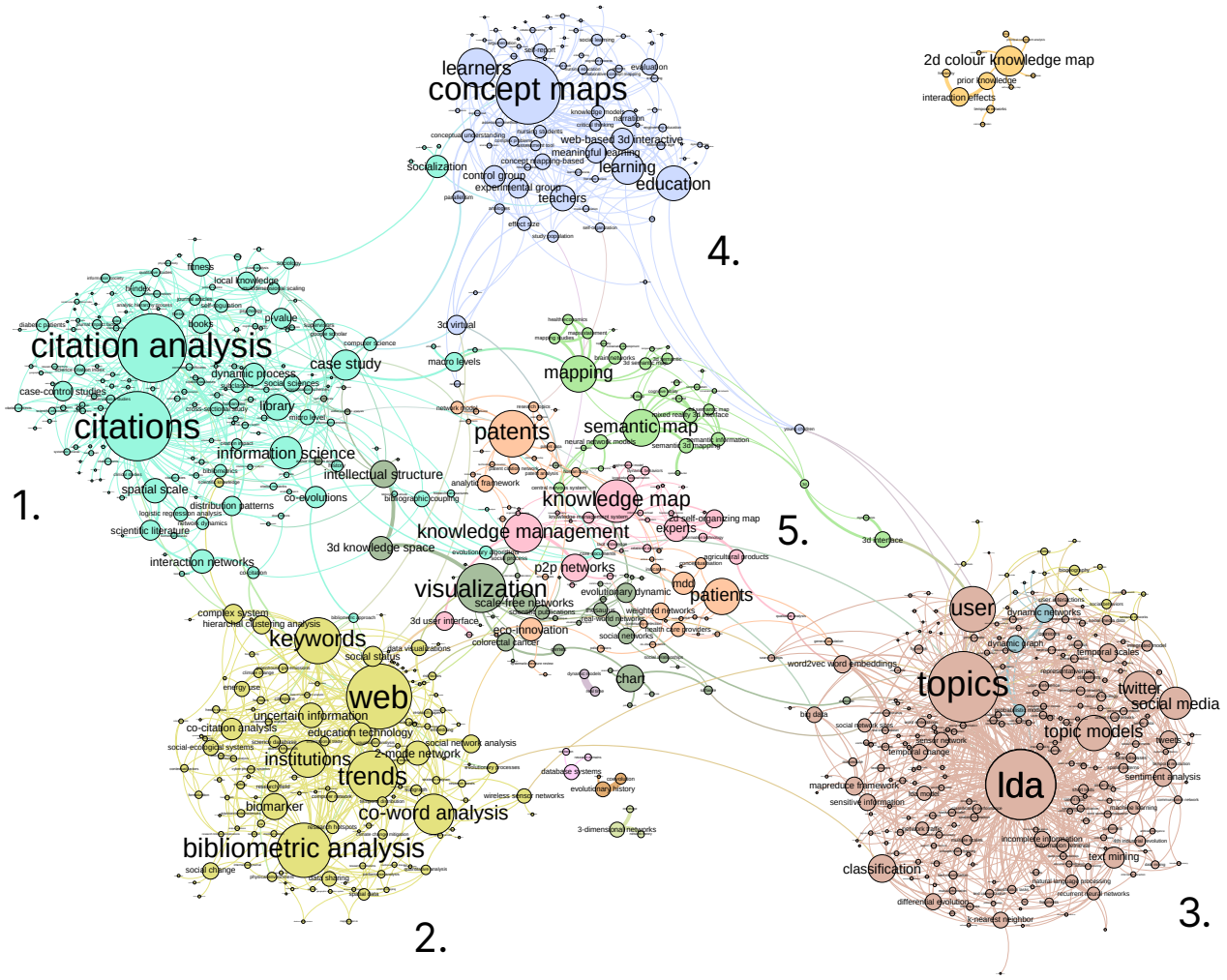

Figure 2. Map of the meta-data of 13844 publications related to *text analysis and knowledge visualization*, extracted from the Web of Science by using a confidence similarity metric. Generated with Gargantext and spatialized with Gephi to highlight scientific communities [1...5]. Interactive version available at http://maps.gargantext.org/maps/sciencemaps/.

2. **Bibliometrics analysis.** In addition to pure citation analysis, the field of *bibliometrics analysis* developed in the early 1970s, in keeping with *bibliographic coupling* and *co-citation* techniques[22,23]. Later on, following the creation of the *Web*, bibliometrics approaches enjoyed a surge of interest with the emergence of *hyperlinked data*[24]. *Visualization analysis* became central since it offered tools to describe *conceptual structures* of science such as *research fronts*, *hot topics* and *trends*, etc.[25–27] – which might also be studied together as socio-semantic networks[28]. *Co-word analysis* is a bottom-up approach. First developed by sociologists in the 1980s[29] to reconstruct the dynamics of *research themes* out of words' *co-occurrence*, it quickly paved the way to hybrid research[30,31]. All these methods have primarily borrowed concepts from *graphs* and *social networks* analysis. The sub-field of *science mapping* aims to explore the *social structures* and *temporal evolution* of *academic research*[32] with the help of computer science techniques[33]. Nowadays, *science maps* are interdisciplinary objects of research resulting from both quali-quantitative and socio-technical processes[34]. The growth of scientific databases has finally stimulated the visualization of wide *citation landscapes*[35] or complex atlases of sciences[36].

3. **Topic modeling.** This field emerged in the early 2000s at the instigation of a community of statisticians who first used the *Latent Dirichlet Allocation* method[37] to characterize *collections of documents*. Although focusing primarily on *document classification*[38], *recommendation*[39] or *sentiment analysis*[40], parts of their most recent works have started to investigate science mapping[41,42].

4. **Concept and semantic maps.** In the 1990s, both mapping techniques gained entry to the broader field of *science of education*[43] as means to support *knowledge integration*[44]. Concept mapping has been deeply influenced by psychology and cognition. A *concept map* can be defined as a *graphical representation* designed to highlight the relationships between ideas or *key concepts*[45]. Its purpose is to clarify a given topic as well as its underlying *cognitive structure*[46,47] by means of *ontologies*, *mind maps*, *mental models*, etc. Unlike co-word approaches,

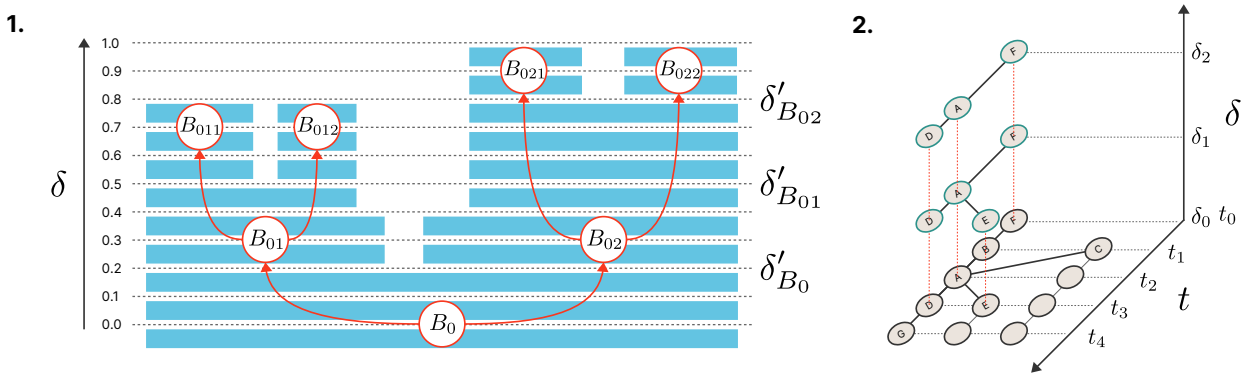

Figure 3. 1. The sea level rise algorithm (the initial branch B_0 breaks into smaller branches $B_{011}, B_{012}, B_{021}, B_{022}$) and 2. the resulting scales of description (a temporal serie of fields $[A...G]$ is associated to each value of δ)

concept maps were initially supposed to translate elements of knowledge issued by *learners* and *teachers* in a top-down way. But the recent influence of data mining methods has reversed this trend by increasing the use of bottom-up recommendation systems or topic detection, along with the introduction of *visualization tools*[48–50].

5. **Domains with peripheral concerns.** Unlike clusters no.1 to no.3, the communities represented by cluster no.5 are not focusing on a single method, but rather borrow existing techniques from the latter to study their own objects of research. Among these peripheral domains, the fields of *knowledge management*[51], *business intelligence*[52] and *patent analysis*[53] stand out.

2.2 Revealing the dynamics of knowledge

Over the last decade, a wide variety of knowledge driven domains such as science have become fertile fields of investigation for the study of time-related dynamics. Being able to access countless digitized archives has acted as an incentive for researchers[54,55]. Each of the communities appearing on Figure 2 has developed its own set of temporally-aware techniques, from citations dynamics[27,54,56,57] to topic modeling over time[42,58–61] and co-word evolution in the case of phylomemies[14,62]. The sub-domain of knowledge dynamics visualization has been investigated alongside and can be characterized by the underlying text-mining techniques, by the nature of the original corpus, by the temporal properties displayed, by the capacity to take multi-level and multi-scale complexities into account and by the interactive features. In Table 1 we identify the main characteristics of our visualization system and compare it to a selection of key papers published during the last decade. By doing so, we review the advantages and disadvantages of existing approaches and highlight the originality of ours.

Except our own visualization system, few previous studies[63–65] have proposed to visualize all the basic temporal properties of phylomemies at once. We even would claim that none of them has investigated the articulation between *level* and *scale* and has tried to foreshadow the multi-level complexity of the visualized phenomenon as we plan to do. For designing the main view of our system, we will used a tree like representation as it seems to be the most convincing way to represent dynamic processes (merge / split events, terms dynamics) regarding the state of the art[64–66].

But contrary to existing approaches[63,64], our system will display more complex inheritance lineages with direct parents / children as well as distant ancestors. Our system will have the particularity to propose an endogenous and hierarchical branches' layout that rests on individual term-to-term temporal relations at a given level (see 3.4) instead of using pre-existing ontologies[67] or top-down topics clustering[65,79]. Finally, we will take advantage of the synchronic multi-scale ranges of description induced by the choice of a given level to fold / unfold the content of a branch without additional calculation (see 3.5) like re-processing tree cuts[65] or still using pre-existing organizations[70].

In a near future, our system will be integrated to the free software Gargantext (see 3.6) to allow for interactions with the original documents directly from the visualization. This final step will close the loop between the raw data, the reconstruction of their latent temporal structures and their interactive exploration. As we attach importance to promote reproducibility and open science, we thus share our source code and experimental data sets (see C). To sum up, we think that even if this paper strictly focuses on phylomemy visualization, the way we tackle the multi-level and multi-scale issues can benefit to the overall domain of knowledge dynamics visualization.

2.2.1 Details of the phylomemy reconstruction process

We review the main features of the phylomemy reconstruction process[14] in B. We now give details of a mechanism recently added[15] to this general process: the *sea level rise algorithm* as it helps to understand the relations between the multi-level and multi-scale properties we aim to visualize.

Sea level rise algorithm. The phylomemy reconstruction process deals with the notions of both *level* and *scale*. In complex systems approaches, levels are generally higher descriptors than scales. Here, we first establishes a *level of observation*, noted λ, namely the level of intrinsic complexity of the point of view on the phylomemy. In the Figure 1, the choice of a level of observation λ comes down to the choice of particular values of inter-temporal similarity, noted δ, that can be different for each field. Mathematically speaking, it is a plaque of the foliation, called thereafter *phylomemetic network*.

Table 1. Comparison between different approaches for the visualization of knowledge dynamics.

paper	this paper	[67]	[68]	[69]	[70]	[71]	[72]	[73]	[63]	[74]	[65]	[66]	[75]	[64]	[76]	[77]	[78]	[79]
publication year	2020	2020	2019	2018	2018	2017	2017	2017	2016	2015	2014	2014	2013	2013	2012	2012	2012	2011
text mining method	CWA	WE & TM	NER	CNA	SA	TM	TM	CA & WE	TM	CNA	TM	CWA	TM	CWA	CNA	CA	CWA & TM	TM
visualization method	tree	map	stream	glyph	stream	stream	map	map	tree & stream	stream	tree & stream	tree	stream	tree	bubble	graph	graph	stream
unstructured text	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓
short text	✓	✓	?	?	?	✗	✗	✓	✓	✓	?	?	?	✓	✓	✓	✗	✓
unconstrained # of fields	✓	✗	✗	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗
evolving fields	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓
re-emerging fields	✓	✗	✓	✗	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✗	?	✓
splitting & merging fields	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✗	✓	✗	✓
adaptive temporal matching	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
multi-level	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
multi-scale	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓
macro-to-micro interactions	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓
endogenous branches' clustering	✓	✗	✗	✓	✓	✓	✓	✗	✓	✗	✓	✗	✓	✗	✗	✓	✗	✓
terms dynamics	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗
faceting	✓	✓	✗	✗	…	✓	…	✓	✓	✗	✓	…	…	?	?	✗	✓	?
linked to the original corpus	…	?	✗	✓	✗	✓	✗	…	✓	✗	?	✗	✗	✗	✓	✗	✓	✗
user's preferences	✓	✓	✗	✗	✗	✓	✓	✓	…	✓	…	✗	✗	…	…	✗	✗	✗
reproducibility	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗
open source	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗

✓: the property is fully part of the study; **…** : the property is not part of the study but could integrate it; ✗: the property is not part of the study or incompatible with the approach; **?** : the elements given in the paper do not allow us to assess this aspect; *text mining method* : Complex Network Analysis (CNA), Co-Word Analysis (CWA), Word Embedding (WE), Topic Modeling (TM), Citation Analysis (CA), Named Entity Recognition (NER), Sentiment Analysis (SA)

The intrinsic complexity of a *phylomemetic network* is described by an objective function F_λ [15] parameterized by λ. This function is based on notions pertaining to information retrieval called *accuracy* and *recall*[‡]. The choice of a level of observation λ is equivalent to the determination, for each field, of the most suitable lineages, *i.e.* the appropriate value for δ, such that the resulting *phylomemetic network* optimizes F_λ.

To find an approximation of this optimal *phylomemetic network* φ_λ, the *sea level rise* algorithm has been proposed [15]. This algorithm starts from the connected components, called *branches* (see B), of the *phylomemetic network* defined at $\delta = 0$. It then gradually and recursively raises the inter-temporal matching threshold δ as explained by Figure 3.1. At a low value of similarity $\delta = 0$, the resulting phylomemetic network φ_0 has few or even a single large connected component that are like wide continents. Later on, the gradual rise of δ splits this land into smaller islands by cutting some of the weakest inter-temporal links. Some branches become siblings while others split up earlier and then evolve separately. By keeping on increasing δ independently within each new branch while $F_\lambda(\varphi)$ is improving, the newly-formed branches eventually drift away from each other, giving shape to the whole phylomemy.

An important consequence of the choice of λ is that it will determine how far we have to go in time to find eligible parents / children. Setting a level of observation has thus a direct impact on the characteristic timescales of the visualized structures. The choice of λ is left to the researcher's discretion in light of her own expertise and research questions, which makes any phylomemy a researcher's perception artifact.

Under the influence of the level of observation λ the reconstruction process produces an endogenous hierarchical clustering of the branches induced by their temporal patterns that defines natural scales of description. For example in the Figure 3.1, B_{011} and B_{012} result from the split of B_{01} which itself comes out of B_0. The scales of description of a branch can be defined by merging its fields according to the different levels of its endogenous hierarchical clustering. For example, from δ_0 to δ_2, the branch displayed in Figure 3.2 can be simplified to the set of fields D, A, F by blending the content of fields C and E in. It is important to note that from one level of observation to another, two scales of description may not be comparable since kinship links might have been reprocessed when moving from one level of observation to the next.

3 Method

After having positioned our contribution regarding the state of the art, we now want to focus on the last step of the reconstruction process $O \rightarrow R \rightarrow V$ and thus describe our visualization system. Once we have reconstructed a phylomemy for a specific level of λ, we extract it (see 3.1) and project it onto a two dimensional space. To that end, we will first establish a *graphical description* of the phylomemy (see 3.2), we will then choose the suitable *axis of navigation* (see 3.3) before designing the *views* (see 3.4) and building the *lenses* of exploration (see 3.5) used by the *Memiescape*: our working demonstrator (see 3.6).

To illustrate our method, we'll use a set of screenshots taken from *Memiescape* and thus visualize the reconstructed phylomemy of the heterogeneous corpus of scientific publications $\mathcal{D}_{cnrs}$ (see Figure 4).

3.1 Pre-processing

Before being visualized through the *Memiescape*, we need to pre-process the phylomemy, i.e. extract a convenient 'slice' from the complex structure of Figure 1. To that end, we start by diving into the branches, so to speak, for extracting the fields with regard to the local evolution of δ. The resulting network is made of branches sorted according to their drifting hierarchy (see Figure 3.1). We then filter this network to remove minor branches, i.e. branches covering less than a minimal number of periods of time. This pruning aims to clarify the future reading of the visualization. Finally, we name the remaining branches by means of a two-terms label. We elect the most frequently emerging term in the targeted branch as the first component of the label. The second one is based on a classical *tf–idf* measure computed within the branch's scope. If a given branch does not contain any emerging term, its label results from the union of the two terms with the highest *tf–idf* score. By doing so, the branch's label should be a reasonable compromise between the specificity and the representativeness of its constitutive vocabulary. In Figure 4.2 for instance, the targeted branch is named after the union of *alzheimer* and *disease* and gathers research focusing on the genetical aspects of this neurodegenerative disease.

3.2 Graphical description

The main elements of a phylomemy are *terms*, *fields* and *branches*. They are subject to the structural constraint: terms $\in$ fields $\in$ branches. In addition, terms and branches evolve through time, from one period to another (forward and backward), by means of *kinship* lines that connect pairs of fields together. These weighted connections result from the inter-temporal matching mechanism and thus convey a similarity score.

But some of the kinship lines might have been cut off by the *sea level rise* algorithm (see 2.2.1). We call these cut-off lines *ghost lines* of the branches' drift. These artifacts are vectors of information: they convey the similarity gap between two consecutive branches namely their hierarchical relation by twos. In addition, we are able to determine the local range of similarity of each branch and deduce their highest value of δ. By using ghost lines and similarity ranges, we can reconstruct the whole drifting history of the phylomemy's branches as a naturally hierarchical process and give insights into the surrounding meta-structure at different levels of observation.

Unlike terms, which might appear over and over again, fields are strictly timestamped within specific branches. But their corresponding dates give us the possibility to enrich each term with dynamical properties that spread along kinship lines or across distant branches. By doing so, we can

[‡] *Accuracy* is the proportion of relevant elements among all the elements retrieved and *recall* is the proportion of relevant elements actually retrieved among all the relevant ones.

determine the terms' frequency of appearance per period or ask whether one of them is emerging or declining:

- **Emerging.** A term is emerging if it appears at a specific period for the first time in the whole phylomemy.
- **Declining.** A term is declining if it is used at a specific period for the last time in the whole phylomemy.

3.3 Axis of navigation

We propose to introduce complementary axes of navigation (foreshadowed in 2.2) by using Saussure's linguistic concepts of *synchrony* and *diachrony*[81]:

- **Synchrony.** Analyzing a language at a particular moment of its history.
- **Diachrony.** Analyzing the historical and temporal evolution of a language.

Applied to our system of visualization, the majority of the elements of a phylomemy will be displayed along a diachronic axis of navigation to highlight their temporal evolution. Yet, some interactive features such as timeless scales of description of the branches will be based on a synchronic axis. We thus think that the combination of these two axis will help us to successfully address the challenge of visualizing both multi-level and multi-scale properties.

3.4 Complementary views

We call *views* a set of dedicated visualizations designed to address different issues but which can still be articulated together as part of a whole system.

3.4.1 The seabed view The first view (Figure 4.1, top part) aims to represent the hierarchical relations between branches and give insights into the multi-level meta-structure of the phylomemy. We call this view the *seabed view* as it extends the metaphor of the *sea level rise* algorithm. The goal here is to project the branches in a two-dimensional topographic space.

We visualize the branches as if they were observed from an elevated point and we use black triangles to symbolize their *peaks* (Figure 4.2). In our system of coordinates, the ordinate goes from 0 to 1 (from top to bottom) and maps the last δ elevation of each branch. The abscissa translates the smallest gap of similarity between two consecutive branches. The seabed view thus helps us to understand a posteriori the outcomes of the sea level rise algorithm and its hierarchical drift: two branches displayed on opposite sides are certain to share no terms. In the Figure 4 for instance, branches related to space exploration, dark matter and galaxies lie on the far left, while branches linked to genome, DNA fragmentation and human cells stay on the far right. In the same way, a branch that ends up at the bottom of the seabed view can be identified as a very specialized branch of knowledge (i.e., it results from a high value of δ).

Finally, we draw a set of isolines around the branches' peaks (Figure 4.2, blue curves) by using their spacial density[§]. The more branches we find side by side, the more isolines we draw. This method is an endogenous way to quickly highlight archipelagos of closely related branches like those focusing on *alzheimer* and *hippocampus* in the Figure 4.2. It gives us a hint about the way the same phylomemy could be shaped under different levels of observation by either foreshadowing what mergers could occur between branches for higher values of λ or giving insights into upcoming splits for lower levels.

3.4.2 The kinship view In the *Origin of Species*[83], Darwin used a single illustration[¶] to suggest that evolution extends along lines throughout time. What he drew was *evolutionary tree* where kinship connections were supposed to leave genetic information from ancestors to future generations. We now propose to use similar trees to represent the full body of our phylomemy's branches and explore the dynamics of knowledge at work. Tree-like visualizations already have a long-established legitimacy, from biology to anthropology[84] as vectors of comparative studies and topographic analysis (see the state of the art 2.1)). Our second view (Figure 4.1, bottom part) is therefore called the *kinship view*.

We have chosen to use the same abscissa coordinates as in the seabed view to sort branches from left to right. We try to maintain their horizontal drifting gap as much as possible (according to scalability constraints, see section 5). Fields (full circles) are then arranged under their respective branches' ticks hanks to the *Graphviz* spatialization algorithm[85]. This algorithm tries to minimize overlapping between fields and intercrossing between links. The size of a field either represents the number of documents it embodies or an arbitrary weight attached by the user to each document of the original corpus. We next appoint the fields' ordinates according to their timestamps: parent fields appear at the top of the kinship view while children fields are at the bottom. Finally, we draw inheritance lines (solid dark lines) between the fields but without arrow: we think that researchers should be free to follow the natural flow of time or go back, up towards the origins of the branches.

Still, a synchronic interpretation remains conceivable by looking horizontally at the fields. For a given year, one can observe a set of contemporary fields distributed among branches whose similarity relationships can be deduced thanks to the seabed view.

3.5 Interacting with the views

Since the late 1960's, the science of graphics has evolved thanks to computer sciences tools to improve data visualization. Nowadays, it is common practice to add interactive mechanisms as a way to enrich the analysis of a given graphical representation[86]. We will now describe three *lenses* of exploration through which to look at the views and interact with them from macro to micro scales.

3.5.1 The macro lens The *macro lens* aims to give an overview of the branches' evolution and relations of similarity. The kinship view includes zoom and drag mechanisms. By zooming in, one can focus on a given field

[§] We use the library https://github.com/d3/d3-contour that relies on a marching squares algorithm[82]

[¶] https://commons.wikimedia.org/wiki/File:Darwin_divergence.jpg

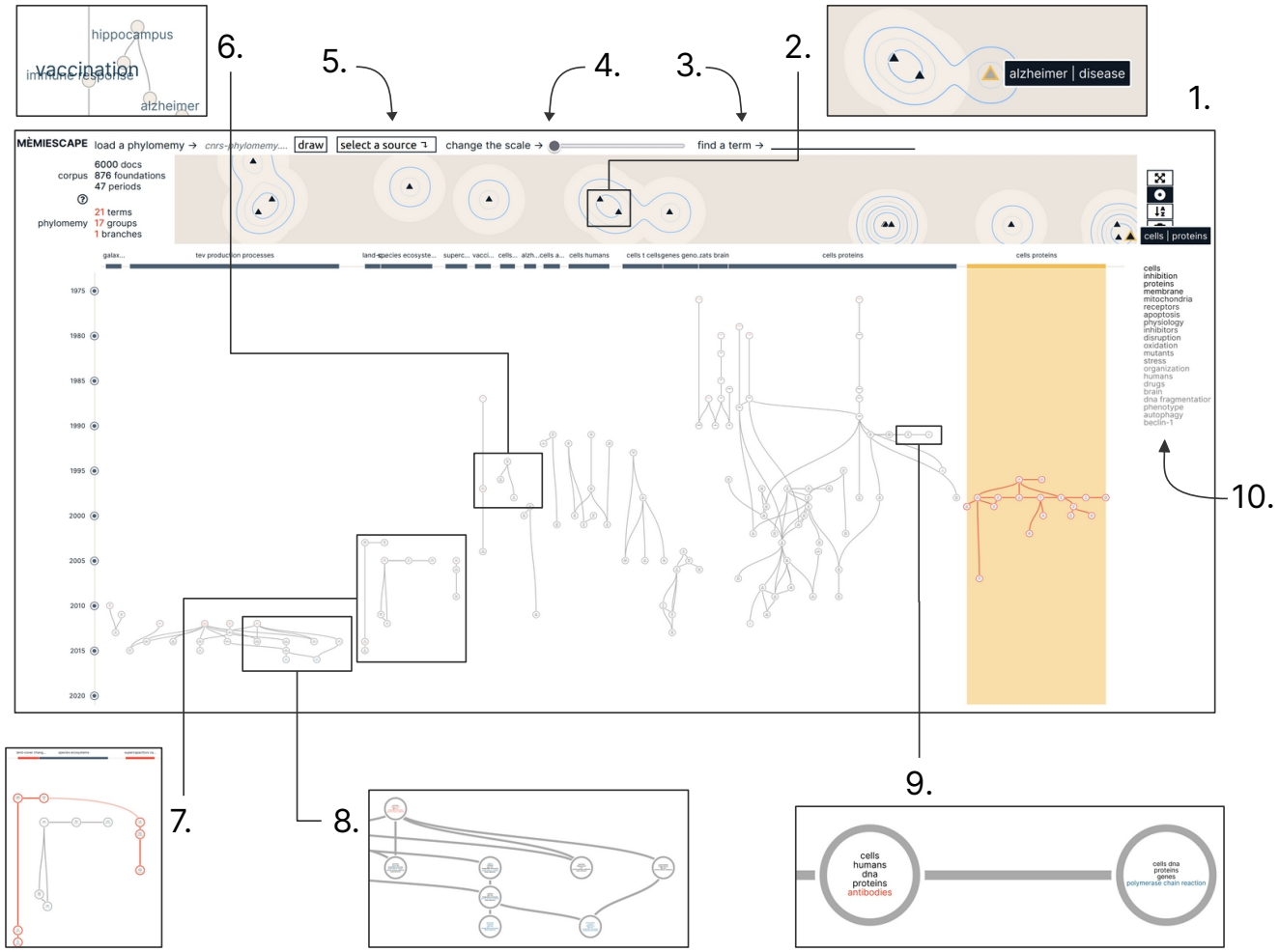

Figure 4. Screenshots taken from the *Memiescape* to visualize the phylomemy of the corpus $\mathcal{D}_{cnrs}$. Inserts $[1...10]$ highlight the main features of the *Memiescape*

and reveal the full name of its branch (Figure 4.8). If the user's cursor moves over a branch's tick, its corresponding fields and peak are highlighted by a yellow thread (Figure 4.1 and Figure 4.2). When a branch is dragged out of the kinship view, its peak switches off in the seabed view – wich acts as a fixed map above the branches to prevent researchers from getting lost in the midst of the phylomemy. When one clicks on a peak, the kinship view is automatically readjusted around the coordinates of the corresponding branch.

3.5.2 The mezzo lens While the *macro lens* outlines the global shape of the phylomemy, *mezzo lens* focuses on its constitutive elements, namely on the emerging terms (see Figure 4.6): these have been extracted beforehand (see 3.1) and are displayed in ordinate according to their date of appearance. The mezzo lens is considered as the landing lens of the Memiescape, as it delivers a reasonable amount of information for exploring the phylomemy at first glance. As for the abscissa, if a term emerges in a single field we obviously reuse that field's coordinates, but if it appears twice or more at the same time, we place it at the barycenter of its emergence fields. An emerging term might also appear written in black or red according to whether it is shared by multiple branches or not. The size of the terms maps their frequency of appearance in the original corpus during the most recent period T_{last}. We also display the full list of terms one might find within a given branch when clicking on its corresponding peak (see Figure 4.10). The opacity of the terms appearing on the lists depends on the number of fields they are part of in the phylomemy. This enables us first to highlight the semantic innovations in each branch (or sets of branches) and therefore the contributions each has made to the whole landscape; and then to show the user the vocabulary still employed at the last stage of the phylomemy. We here make the assumption that the point of view of the user is situated in time, as one usually tries to understand the current state of a given element of knowledge at T_{last} regarding its historical evolution since T_{first} (e.g., while creating a bibliography). Figure 4.6 here illustrates the use of the mezzo lens to reveal major breakthroughs in the research conducted by the CNRS on *immune response* and *vaccination* between 1987 and 2004. By clicking on one of these terms, the user can switch to the *micro lens*.

3.5.3 The micro lens The *micro lens* is designed to dive fully into the textual content of the phylomemy. It first displays terms within their respective fields (Figure 4.8) before outlining the declining and emerging ones with a color code: blue for declining terms (e.g., *polymerase chain reaction* in Figure 4.9) and red for emerging terms (e.g., *antibodies* in Figure 4.9). When the user clicks on a term or used the search box (Figure 4.3), our phylomemy

reveals the way this term spreads among and across the branches (Figure 4.7). We put all the kinship lines linking together fields containing the targeted term in red and draw additional light red lines between any distant branches that might have been using it beforehand. In Figure 4.4, we thus highlight the shared use of the term *carbon* from the branch *land-cover change* (2000's - 2010's) to the branch *supercapacitors* (2000's). A term's click also triggers an update of the related terms list (Figure 4.10) by displaying all its corresponding co-used terms. We also bring the related branches to the front of the seabed view and add a *find more* link to the Wikipedia's page of the targeted term if it exists. The micro lens thus makes it possible to follow the internal dynamics of the phylomemy and to understand trans-disciplinary influences through semantic dissemination between branches. In addition, we allow the user to go trough faceted search: by selecting a specific *source* (Figure 4.5) we point out with a dedicated color all the fields concerned by this source. We generically call *source* a countable and discriminating attribute added to each original document like the journal's names in the case of a corpus of scientific publications.

3.5.4 Interactive scales of description As foreshadowed in 2.2.1, the choice of a level of observation λ induces within each branch a range of scales of description we now want to use for interacting with the phylomemy.

To that end, we collect all the kinship links of a given branch before sorting and grouping them by ascending value of δ similarity. We thus create a range of scales of description that defines a family of hierarchical clustering tasks. By selecting a branch in our system, we unlock a slider (Figure 4.10) configured on the fly with the values of the corresponding scales' range. When the user goes over a given scale, we prune the kinship links whose value of δ similarity is lower than this scale. We then synchronically merge the semantic content of the fields that would have belonged to the same connected component if this kinship links had not existed. We thus endogenously cluster the fields that belong to the same period of time without additional calculation. The whole process takes care of not breaking the continuity of the branch. By doing so, we propose to discover each branch from its finest scale of description to its most aggregated and pruned one regarding a chosen level of observation as illustrated by the Figure 5.

This articulation between the temporal properties of the current level of observation and the resulting scales of description (in the kinship view) together with the insights into the surrounding higher / lower levels of complexity (in the sea bed view) contribute to the originality of our visualization system. Altogether embodied within the Memiescape, these mechanisms help us to explore and interact with the complex information initially hidden in the Figure 1.

3.6 Implementation

In terms of technical support, the free software *Gargantext* provides us with a set of fully implemented functions for the reconstruction of phylomemies. We include the most recent research developments[15] on phylomemetic projections extraction* (see 3.1) and export them as pre-spatialized Json files by means of *Graphviz*. We then load those files within *Memiescape*, our dedicated demonstrator for the visualization of phylomemies.

Gargantext. It is a free text-mining software† developed in *Haskell*. Gargantext makes it possible to turn knowledge structures into tangible artifacts[19]. Gargantext addresses, by design, the user's role in knowledge-mining tasks and therefore incorporates collaborative, cumulative and collective features. Semantic maps are created thanks to real-time peer collaboration through visualizations, easy reuse of former materials and machine learning on individual and collective past usages. The revealed shapes are consequently the outcomes of a series of reflexive choices and cumulative expertise. By using Gargantext, we aim to guarantee the easy reproducibility of our results and shorten 'time-to-innovation' cycles.

Memiescape. It is a standalone Web demonstrator usable in a wide number of scenarios without online dependency‡. Because of scalability concerns, almost all text-mining aspects are done upstream within Gargantext. The remaining tasks are processed in the browser with JavaScript and React elements of codes to manage the views in real time. Graphics and interactive mechanisms are made in pure d3js. Memiescape is published under Gargantext licences: aGPLV3 and CECILL variant Affero compliant§.

4 Results

We will now go through three different use cases. Each of them pursues a specific goal: with 4.1 we put the Memiescape into practice and illustrate the way we can navigate through the temporal information hidden within a corpus of scientific publications; with 4.2 we externally validate our system of visualization by reconstructing from a diachronic perspective the whole state of the art reviewed in 2.1; with 4.3 we report a user experiment in which a group of epidemiologists were confronted with the visualization of the worldwide dynamics of Covid-19 vaccines researches. We rely for this on manually annotated screenshots taken from Memiescape and summarizing live explorations.

4.1 An interdisciplinary corpus of academic publications

In section 3, we illustrated our technical choices through the visualization of the corpus $\mathcal{D}_{cnrs}$. Interdisciplinary by nature, this corpus gathers some of the most internationally influential publications (extracted from the WoS) recently carried out by researchers affiliated to the french CNRS. Such a collection could be used as a means to understand the dynamics of research and innovation at a national

*Code is available at https://gitlab.iscpif.fr/gargantext/haskell-gargantext/tree/master

†See https://gargantext.org, https://www.haskell.org/, https://reactjs.org/ & https://d3js.org/

‡Code is available at https://gitlab.iscpif.fr/qlobbe/memiescape/tree/v2

§See https://gitlab.iscpif.fr/humanities/gargantext/blob/stable/LICENSE

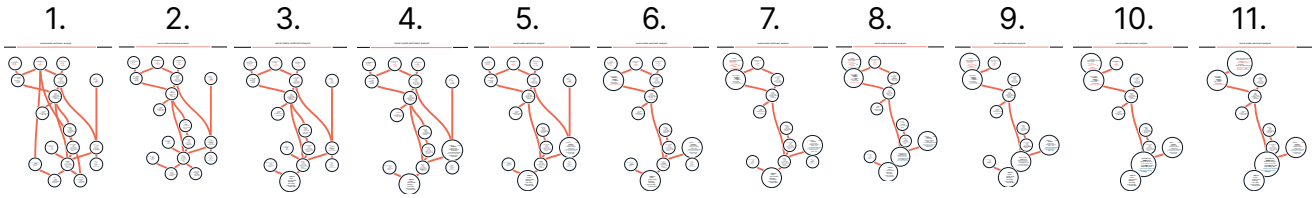

Figure 5. The evolution of the branch *social media and sentiment analysis* extracted from the phylomemy $\mathcal{D}_{maps}$ (see Figure 7) for 11 scales of description (from 1. for the highest scale to 11. for the lowest scale) at level 0.5

Figure 6. Manually annotated phylomemy of the corpus $\mathcal{D}_{cnrs}$, red rectangles $[1...9]$ highlight noteworthy branches of knowledge. Interactive version available at http://maps.gargantext.org/phylo/cnrs_top_cited/memiescape/

scale. The collective landscape shaped by this corpus may be of interest for historians, sociologists or philosophers who investigate the underlying mechanisms of Science: academic collaboration networks, national policy effects, funding trends, etc. Let us then go through a more detailed exploration of this corpus, using both its phylomemy (see Figure 6) and the original publications.

The reconstructed phylomemy of $\mathcal{D}_{cnrs}$ shows modern science's global tendency to focus on the microscopic world as a mean to understand larger natural phenomenons – from human health to biosphere changes and universe expansion. In the late 1980's for instance, neuroscience paved the way to the comprehension of *brain illness* (branches no.1) by first diving into the nervous system of rats before investigating the role of the *hippocampus* in *memory consolidation* processes and *alzheimer's disease* (branches no.2). Later on, biologists built on 1990's genomics improvements (branches no.3) to grasp how *mitochondria* influenced *nuclear apoptosis* mechanisms, i.e. *cells death* and *cancers* (branches no.4). During the 2000's ans 2010's, genetics has led to the decoding of full *genome sequences* (branches no.5) which were then used in the characterization of *species* (plants, bacteria, etc.) or for treating *genetic human diseases*. At the same time, medicine and pathology started to make use of genomics as well in an attempt to improve our adaptive immune system against viruses (branches no.6). Cloning techniques like *monoclonal antibody* have here been applied to prevent autoimmune diseases or induce immune responses against targeted cancer cells. As for environmental researchers (branches no.7), they started to push *global warming* to the fore of ecology concerns at the turn of the 2000's. They pointed out the degradation of *carbon exchanges* between *oceans*, *tropical forests* and the whole biosphere as well as an increasing *loss of biodiversity*. Beyond Earth and its atmospheric concentrations of C02, astrophysicists then tracked *molecular gas* (like *carbon monoxide*) and *cosmic dust* (branches no.8) to discover *galaxies* inside the Hubble deep field or to follow their evolution from *star-forming* galaxies to mergers. Nowadays, some of the most influential CNRS publications come from the use of the *large hadron collider* (branches no.9), a *particle accelerator* involved in the discovery of the *Higgs boson* and designed to test theoretical predictions in the fields of particle physics.

4.2 The historical evolution of text analysis and knowledge visualization

In section 2.1, we used a semantic map generated by Gargantext (see Figure 2) to position our approach. We will now go over the reconstruction of its corresponding

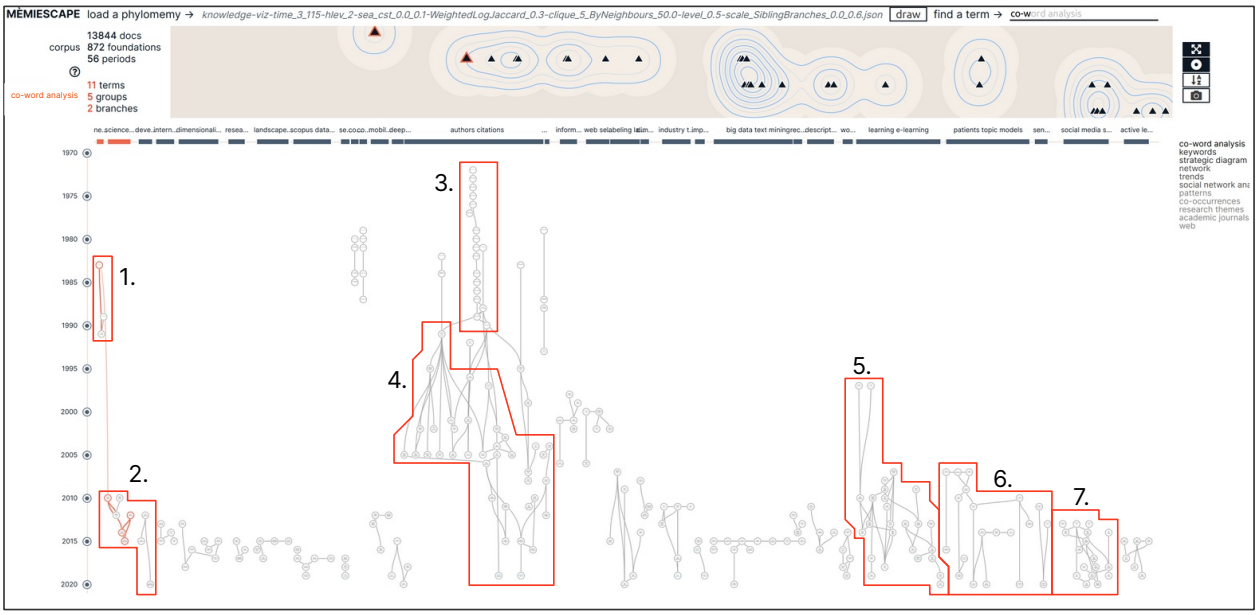

Figure 7. Manually annotated phylomemy of the $\mathcal{D}_{maps}$ corpus at level 0.5, red rectangles $[1...7]$ highlight noteworthy branches, red curves reveal the spread of *co-word* approaches across the branches. Interactive version available at http://maps.gargantext.org/unpublished_maps_phylo/knowledge_visualization/memiescape/

phylomemy and validate it in light of the state of the art detailed in section 2.1. We will also add a few temporal observations.

Figure 7 first outlines the evolution of *co-occurrence* and *co-word* analyses [17,89]. These were applied in the late 1980's (branch no.1) to study paired data within a given collection of documents and, more specifically, pairs of terms for *co-word* approaches. Both paradigms then enjoyed a revival of interest in the mid-2000's (branches no.2) as a result of the ICT revolution. They aimed to reveal the structural and dynamical evolution of elements of knowledge by focusing on temporal trends as well as paradigm shifts in science and research fronts [14,15,18]. Our phylomemies are, in a way, heirs to these paradigms. Figure 7 also shows that the classical field of *citation analysis* was predominant during the 1970's (branches no.3) before passing the baton to what would become the core of *bibliometry* and *scientometry* in the early 1990's (branches no.4); who in turn took advantage of the emergence of large *scientific databases* and new *web* resources to investigate the fields of *co-citation analysis* and *bibliometric indices*. In the 2000's, *information retrieval* techniques started to be actively used and, at the same time, the long-established field of *concept mapping* found concrete applications in the domains of *education* and *learning process* (branches no.5). A few years later, *topic modeling* rose and quickly disseminated accross various scientific fields (branches no.6), from *patents* analysis to *recommendations* systems and exploration of *social media footprints* (branches no.5).

4.3 The tracks of Covid-19 vaccines

With this last casework, we aim to reconstruct the evolution of 1263 timestamped descriptions of clinical trials related to the Covid-19 vaccines and published between March 2020 and June 2021. Here, medical descriptions have been dated according to their week of deposit within the *Covid-19 WHO database* and curated by a consortium of epidemiologists [90]*. This database has been later compiled into a textual corpus named $\mathcal{D}_{ct}$. During the reconstruction process we have chosen to match the size of the resulting fields with the total number of people who were recruited to try the corresponding vaccine candidates.

As a users experiment, we have confronted the visualization of the Figure 8 to a group of epidemiologists. By using the Memiescape, the latter have quickly identified the five main research tracks followed to discover the Covid-19 vaccines: *non covid vaccine*, *protein subunit*, *inactivated virus*, *non replicating viral vector* and *rna based*. Within the branch *rna based*, they have retrieved the sub-paths that led to the release of both *Pfizer-BioNTech* (Figure 8.1 see the path of the *bnt162b2* rna vaccine) and *Astrazeneca* vaccines (Figure 8.2 see the path of the adenovirus *chadox1*). The existence of this branch with two clear sub-paths corresponding to very different approaches immediately draw the attention of the epidemiologists and an answer to a question they were recently asked appeared to them immediately: Has a combination between rna vaccines and non rna vaccines been already tested? The answer was at the intersection of the two sub-paths (Figure 8.3) and was responsible for the partial late merging of the two sub-branches between March and April 2021 with a clinical trial testing both *bnt162b2* and *chadox1*.

We have also provided the epidemiologists with four different faceting features: filtering the fields by trial phases, by funding, by involved countries and by the presence of an associated publication or not. The epidemiologists have found the last one to be of particular interest as it translates the very making of science as an ongoing process. The

*The original database can be downloaded here https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.

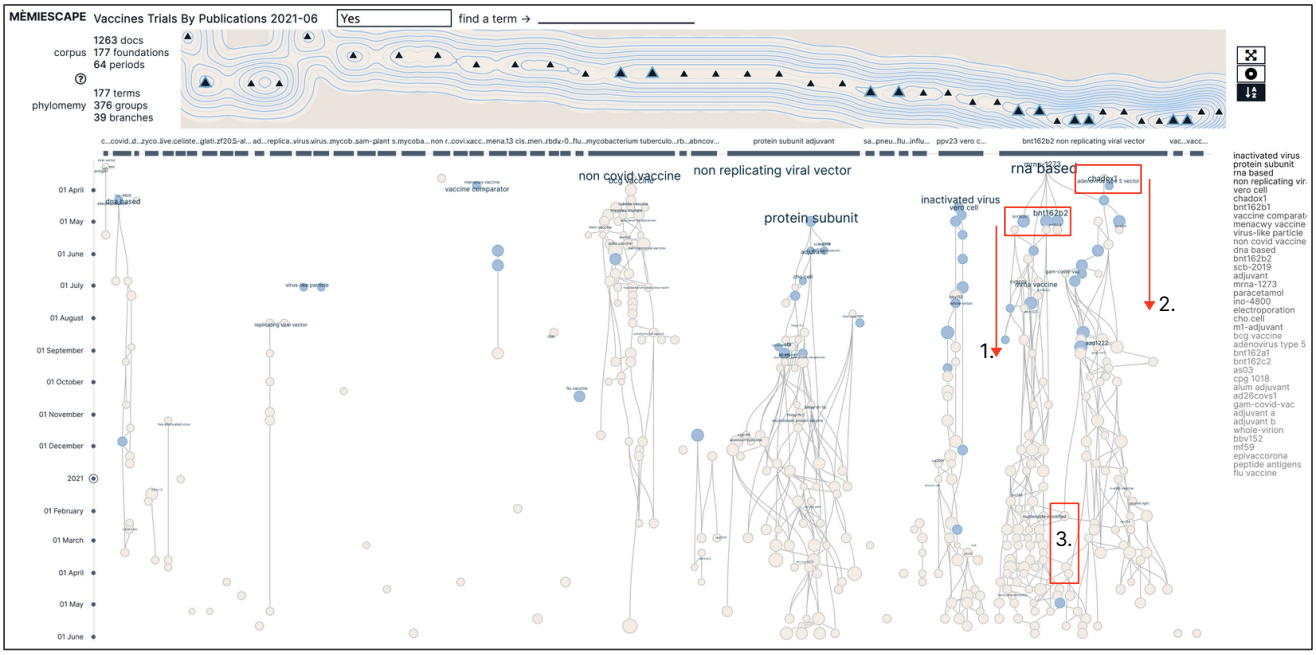

Figure 8. Manually annotated phylomemy of the $\mathcal{D}_{ct}$ corpus. The landing view highlights the four main research tracks followed to discover the Covid-19 vaccines between March 2020 and June 2021. The sub-branches *bnt162b2 rna* (1.) and *chadox1* (2.) evolve apart for months before merging in March 2021 (3.). Interactive version available at http://maps.gargantext.org/unpublished_maps_phylo/vaccines_publications/

epidemiologists could thus see immediately that the branch *non covid vaccine* was intensely explored at the beginning of the pandemic before being slowly abandoned in light of trials results with almost no publications , meaning that no significant discoveries were made in this branch.

To sum up, this phylomemy bears witness of a worldwide effort to find an effective vaccine. It is our belief that the visualization of phylomemies could be a powerful tool to foster collective coordination between researchers.

5 Discussion

5.1 Wrapping-up a coherent multi-scale and multi-level methodology of exploration

We have to admit that screenshots taken from Figure 6, 7 and 8 fail to translate the way we actually navigate through a phylomemy. Future improvements should therefore address the question of how to effectively translate the outcomes of an exploration in a static illustration. Yet, we think that what interactions between our views (see 3.4) and lenses (see 3.5) already exist are worthwhile foundations for a convincing methodology of exploration. Users in particular have emphasized in their feedback how phylomemies' multi-level and multi-scale properties stimulated their curiosity and made them want to dive deeper.

We have been inspired by the science of complex systems, in which the researcher can switch between *micro*, *mezzo*, and *macro* scales. Here, the macro lens (see Figure 4.1) gives an overview of the temporal evolution, the mezzo lens (see Figure 4.3) helps to characterize the branches and the micro lens (see Figure 4.4) reveals the underlying semantic structure. Natural systems and phenomenons are indeed often composed of elements interacting from one scale to another. Different elements might be relevant for different scales and, for instance, micro relationships (e.g. terms used in the same document) might induce the emergence of macro structures.

Last but not least, distinct phylomemies reconstructed at different levels of observation λ (see 2.2.1) can be jointly explored in *Memiescape* to reveal the whole spectrum of specialization covered by a research domain[15], as illustrated by the Figure 9 and thus fully unfold the structure of Figure 1.

5.2 Scalability issues

- *computational scalability*: The algorithmic complexity of the phylomemy reconstruction workflow proposed in Chavalarias et al. (2020)[15] is more or less linear regarding the number of documents but depends heavily on the size of the list of terms upon which the phylomemy is reconstructed (see 2.2.1) and on the clustering algorithm chosen to define the fields. In the example chosen in this paper, fields are defined as maximal cliques, which worst-case time complexity is $O(3^{\frac{n}{3}})$ for an n-vertex graph[92]. Hopefully, semantic networks are generally sparse such that the maximal cliques algorithm has always been tractable in reasonable time.

  As for our visualization concerns, we haven't identified any complexity issue so far. But we have to notice some limitations in our current implementation: when the Memiescape tries to visualize a phylomemy greater than 1000 groups, the *Firefox* browser slows down and struggles to display all the graphical elements. On that particular point, we are confident that future technical developments might improve the capacity of the Memiescape.

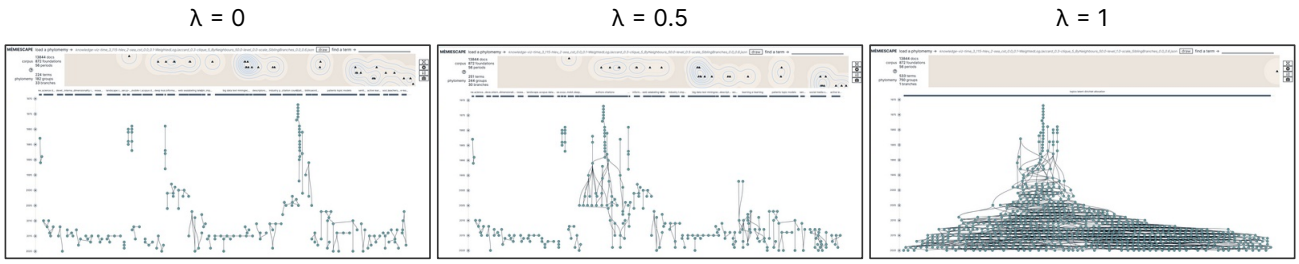

Figure 9. The phylomemy of the corpus $\mathcal{D}_{maps}$ reconstructed for three different levels of observation such as $\lambda \in [0, 0.5, 1]$

- *graphical scalability*: in Figure 6 or Figure 7, we've noticed that large branches' names sometime overlapped in the kinship view. But the name-shortening mechanism already implemented does not totally solve this issue. As a possible answer, we might later propose a zoom technique based on the importance of the branches or try to develop a non overlapping spatialization method for names and texts.

6 Conclusion

A phylomemy is a complex object that winds up in a high dimensional space[15]. The originality of our contribution has been to propose a visualization system that translates in graphical and interactive ways the multi-level and multi-scale properties of a phylomemy (see 3). Our approach thus enriches the state of the art of knowledge dynamics visualization by introducing the possibility to navigate through the same corpus of documents at different levels of observation and to explore it by means of various scales of description (see 3). We have then reviewed three use cases of our system: an illustrative application, a validation and a users experiment (see 4). Our system has also been fully implemented as a free software: the Memiescape (all the use cases are accessible online) and will soon be integrated to the collaborative text-mining platform Gargantext to ensure the continuity between the original corpus of documents and its final representation. By doing so, we should in time be able to create a flowing link connecting any term of the kinship view with its corresponding timestamped publication and thus complete our multi-level and multi-scale methodology of exploration (see 5).

With that in mind, we think that merging our visualization approach with the reflexive and collaborative features of Gargantext[19] will make the exploration become an active process and one that allows researchers to experience the tangible nature of textual data. Since we consider phylomemies as artifacts for the researcher's perception (see 2.2.1), we want to give him/her even more of a central role, set him/her in motion among the original corpora, project him/her through the whole reconstruction process. Future works will therefore be dedicated to the investigation of *tangible exploratory data analysis*: a new 'doorway' methodology for the exploration and visualization of the hidden structure and dynamics of knowledge. This notion will question both the current and the upcoming shapes of a phylomemy thanks to a continuum of iterative loops of analysis. What is the nature of the semantic landscape I'm browsing through? What collection of documents could be missing? What is hidden beyond the borders of my corpus? What new branch of knowledge could appear if I enrich this lineage with an other? What innovative concept could emerge in a near future? Embedded within Gargantext, phylomemies will become open playgrounds where researchers are free to experience multiple round trips from the constitution of their corpora to the collaborative annotation of their visualizations.

References

1. Stiegler B. Leroi-gourhan: l'inorganique organisé. *Les Cahiers de médiologie* 1998; (2): 187–194.
2. Febvre L and Martin HJ. *L'apparition du livre*. Albin Michel, 2013.
3. Borgman CL. *From Gutenberg to the global information infrastructure: access to information in the networked world*. Mit Press, 2003.
4. Lobbé Q. Where the dead blogs are. In *International Conference on Asian Digital Libraries*. Springer, pp. 112–123.
5. Rogers R. *Digital methods*. MIT press, 2013.
6. Manovich L. The science of culture? social computing, digital humanities and cultural analytics 2015; .
7. d'Alembert JLR. *Discours préliminaire de l'encyclopédie: publié intégralement d'après l'édition de 1763 avec les avertissements de 1759 et 1763, la dédicace de 1751, des variantes, des notes, une analyse et une introduction*. A. Colin et cie, 1894.
8. Manfroid S. *Paul Otlet, fondateur du Mundaneum (1868-1944): Architecte du savoir, artisan de paix*. les Impressions nouvelles, 2010.
9. Chavalarias D. Formes collectives. *Le Genre humain* 2019; (1): 145–152.
10. Bonabeau E and Theraulaz G. *Intelligence collective*. Hermes Paris, France, 1994. Bibtex: bonabeauIntelligence1994.
11. Theraulaz G and Bonabeau E. A brief history of stigmergy. *Artificial life* 1999; 5(2): 97–116.
12. Lee K, Lee J, Kim D et al. Controversy visualization : How controversial public discourse in wikipedia articles evolves over time. *Archives of Design Research* 2017; 30: 57–69. DOI: 10.15187/adr.2017.11.30.4.57.
13. Bourgine P, Brodu N, Deffuant G et al. Formal epistemology, experimentation, machine learning. In *HAL Archives Ouvertes*. https://hal.archives-ouvertes.fr/hal-00392486, Chavalarias et al. ISBN https://hal.archives-ouvertes.fr/hal-00392486, 2009. pp. 10–14.
14. Chavalarias D and Cointet JP. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one* 2013; 8(2).

15. Chavalarias D, Lobbé Q and Delanoë A. Draw me science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies forthcoming; .
16. Kitchin R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE, 2014. ISBN 978-1-4739-0826-0. Google-Books-ID: GfOICwAAQBAJ.
17. Callon M, Courtial JP and Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics* 1991; 22(1): 155–205.
18. Delanoe A, Chavalarias D and Anglade A. Dematerialization and environment: a text-mining landscape on academic, blog and press publications. In *ICT for Sustainability 2014 (ICT4S-14)*. Atlantis Press.
19. Delanoë A and Chavalarias D. Mining the digital society - Gargantext, a macroscope for collaborative analysis and exploration of textual corpora. *Forthcoming* 2020; .
20. Lobbé Q. Replication data : Exploring, browsing and interacting with multi-scale structures of knowledge, 2021. DOI:10.7910/DVN/SLARHQ. URL https://doi.org/10.7910/DVN/SLARHQ.
21. Garfield E. Citation analysis as a tool in journal evaluation. *Science* 1972; 178(4060): 471–479.
22. Kessler M. Bibliographic Coupling Between Scientific Papers. *American Documentation* 1963; 14(1): 10–&. DOI:10.1002/asi.5090140103. 01294 WOS:A19632554A00006 bibtex: kesslerBibliographic1963.
23. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science* 1973; 24(4): 265–269.
24. Kleinberg JM. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 1999; 46(5): 604–632. Publisher: ACM New York, NY, USA.
25. White HD and McCain KW. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science* 1998; 49(4): 327–355. DOI:10.1002/(SICI)1097-4571(19980401)49:4⟨327::AID-ASI4⟩3.0.CO;2-W. URL http://doi.wiley.com/10.1002/%28SICI%291097-4571%2819980401%2949%3A4%3C327%3A%3AAID-ASI4%3E3.0.CO%3B2-W.
26. Börner K, Chen CM and Boyack KW. Visualizing knowledge domains. *Annual Review of Information Science and Technology* 2003; 37: 179–255. DOI:10.1002/aris.1440370106. 00868 WOS:000179918000006 bibtex: bornerVisualizing2003.
27. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 2006; 57(3): 359–377. DOI:10.1002/asi.20317. URL http://doi.wiley.com/10.1002/asi.20317.
28. Roth C and Cointet JP. Social and semantic coevolution in knowledge networks. *Social Networks* 2010; 32(1): 16–29. DOI:10.1016/j.socnet.2009.04.005. URL http://www.sciencedirect.com/science/article/pii/S0378873309000215. Bibtex: rothSocial2010.
29. Callon M, Rip A and Law J. *Mapping the dynamics of science and technology: Sociology of science in the real world*. Springer, 1986.
30. Braam RR, Moed HF and van Raan AFJ. Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science* 1991; 42(4): 233–251. DOI:10.1002/(SICI)1097-4571(199105)42:4⟨233::AID-ASI1⟩3.0.CO;2-I. URL http://doi.wiley.com/10.1002/%28SICI%291097-4571%28199105%2942%3A4%3C233%3A%3AAID-ASI1%3E3.0.CO%3B2-I. 00352 bibtex: braamMapping1991.
31. Boyack KW and Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology* 2010; 61(12): 2389–2404. DOI:10.1002/asi.21419. URL http://doi.wiley.com/10.1002/asi.21419.
32. Small H. Visualizing science by citation mapping. *Journal of the American society for Information Science* 1999; 50(9): 799–813.
33. Harries G, Wilkinson D, Price L et al. Hyperlinks as a data source for science mapping. *Journal of Information Science* 2004; 30(5): 436–447.
34. Chen C and Song M. Visualizing a field of research: A methodology of systematic scientometric reviews. *PloS one* 2019; 14(10).
35. Small H. Update on science mapping: Creating large document spaces. *Scientometrics* 1997; 38(2): 275–293.
36. Brner K. *Atlas of science: Visualizing what we know*. The MIT Press, 2010.
37. Blei DM, Ng AY and Jordan MI. Latent dirichlet allocation. *Journal of machine Learning research* 2003; 3(Jan): 993–1022.
38. Wei X and Croft WB. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. pp. 178–185.
39. Wang C and Blei DM. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. San Diego, California, USA: ACM Press. ISBN 978-1-4503-0813-7, p. 448. DOI:10.1145/2020408.2020480. URL http://dl.acm.org/citation.cfm?doid=2020408.2020480.
40. Lin C and He Y. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. pp. 375–384.
41. Millar JR, Peterson GL and Mendenhall MJ. Document clustering and visualization with latent dirichlet allocation and self-organizing maps. In *Twenty-Second International FLAIRS Conference*.
42. Yang Y, Yao Q and Qu H. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics* 2017; 1(1): 40–47.
43. Novak JD. Concept mapping: A useful tool for science education. *Journal of research in science teaching* 1990; 27(10): 937–949.
44. Jonassen DH, Reeves TC, Hong N et al. Concept mapping as cognitive learning and assessment tools. *Journal of interactive learning research* 1997; 8(3): 289.
45. Kinchin IM. Concept mapping in biology. *Journal of biological education* 2000; 34(2): 61–68.

46. Rohrer R, Ebert D and Sibert J. The shape of shakespeare: visualizing text using implicit surfaces. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*. pp. 121–129, 160. DOI:10.1109/INFVIS.1998.729568.
47. Benevene P, Kong E, Barbieri B et al. Representation of intellectual capital's components amongst italian social enterprises. *Journal of Intellectual Capital* 2017; .
48. Nesbit JC and Adesope OO. Learning with concept and knowledge maps: A meta-analysis. *Review of educational research* 2006; 76(3): 413–448.
49. Romero C and Ventura S. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications* 2007; 33(1): 135–146.
50. Slater S, Joksimović S, Kovanovic V et al. Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics* 2017; 42(1): 85–106.
51. Lin Fr and Hsueh Cm. Knowledge map creation and maintenance for virtual communities of practice. *Information processing & management* 2006; 42(2): 551–568.
52. Pyo S. Knowledge map for tourist destinations—needs and implications. *Tourism Management* 2005; 26(4): 583–594.
53. Tseng YH, Lin CJ and Lin YI. Text mining techniques for patent analysis. *Information processing & management* 2007; 43(5): 1216–1247.
54. Chen C, Chen Y, Horowitz M et al. Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics* 2009; 3(3): 191–209. DOI:10.1016/j.joi.2009.03.004. URL http://www.sciencedirect.com/science/article/pii/S1751157709000236.
55. Zeng A, Shen Z, Zhou J et al. The science of science: From the perspective of complex systems. *Physics Reports* 2017; 714-715: 1–73. DOI:10.1016/j.physrep.2017.10.001. URL http://www.sciencedirect.com/science/article/pii/S0370157317303289.
56. Ramos-Rodríguez AR and Ruíz-Navarro J. Changes in the intellectual structure of strategic management research: a bibliometric study of theStrategic Management Journal, 1980–2000. *Strategic Management Journal* 2004; 25(10): 981–1004. DOI:10.1002/smj.397. URL http://doi.wiley.com/10.1002/smj.397.
57. Rosvall M and Bergstrom CT. Mapping Change in Large Networks. *PLOS ONE* 2010; 5(1): e8694. DOI:10.1371/journal.pone.0008694. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008694.
58. Wang X and McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 424–433.
59. Wang C, Blei D and Heckerman D. Continuous Time Dynamic Topic Models. *arXiv:12063298 [cs, stat]* 2015; URL http://arxiv.org/abs/1206.3298. ArXiv: 1206.3298.
60. Cui W, Liu S, Wu Z et al. How Hierarchical Topics Evolve in Large Text Corpora. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 2281–2290. DOI:10.1109/TVCG.2014.2346433. URL http://ieeexplore.ieee.org/document/6875938/.
61. Minjeong K and et al. TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections, 2017.
62. Rule A, Cointet JP and Bearman PS. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences* 2015; : 201512221DOI:10.1073/pnas.1512221112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1512221112. 00000.
63. Liu S, Yin J, Wang X et al. Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics* 2016; 22(11): 2451–2466. DOI:10.1109/TVCG.2015.2509990.
64. Shahaf D, Yang J, Suen C et al. Information cartography: Creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13, New York, NY, USA: Association for Computing Machinery. ISBN 9781450321747, p. 1097–1105. DOI:10.1145/2487575.2487690. URL https://doi.org/10.1145/2487575.2487690.
65. Cui W, Liu S, Wu Z et al. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 2281–2290. DOI:10.1109/TVCG.2014.2346433.
66. Greene D, Archambault D, Belák V et al. Textluas: tracking and visualizing document and term clusters in dynamic text data. *arXiv preprint arXiv:150204609* 2014; .
67. Li Z, Zhang C, Jia S et al. Galex: Exploring the evolution and intersection of disciplines. *IEEE Transactions on Visualization and Computer Graphics* 2020; 26(1): 1182–1192. DOI:10.1109/TVCG.2019.2934667.
68. Dang T, Nguyen HN and Pham V. WordStream: Interactive Visualization for Topic Evolution. In Johansson J, Sadlo F and Marai GE (eds.) *EuroVis 2019 - Short Papers*. The Eurographics Association. ISBN 978-3-03868-090-1. DOI:10.2312/evs.20191178.
69. Perez-Messina I, Gutierrez C and Graells-Garrido E. Organic visualization of document evolution. In *23rd International Conference on Intelligent User Interfaces*. IUI '18, New York, NY, USA: Association for Computing Machinery. ISBN 9781450349451, p. 497–501. DOI:10.1145/3172944.3173004. URL https://doi.org/10.1145/3172944.3173004.
70. Cuenca E, Sallaberry A, Wang FY et al. Multistream: A multiresolution streamgraph approach to explore hierarchical time series. *IEEE Transactions on Visualization and Computer Graphics* 2018; 24(12): 3160–3173. DOI:10.1109/TVCG.2018.2796591.
71. Yang Y, Yao Q and Qu H. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics* 2017; 1(1): 40–47. DOI:https://doi.org/10.1016/j.visinf.2017.01.005. URL https://www.sciencedirect.com/science/article/pii/S2468502X17300074.
72. Kim M, Kang K, Park D et al. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(1): 151–160. DOI:10.1109/TVCG.2016.2598445.
73. Berger M, McDonough K and Seversky LM. cite2vec: Citation-driven document exploration via word embeddings. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(01): 691–700. DOI:10.1109/TVCG.2016.2598667.
74. Weiler A, Grossniklaus M and Scholl MH. The stor-e-motion visualization for topic evolution tracking in text data streams. In *Proceedings of the 6th International Conference*

on Information Visualization Theory and Applications - IVAPP, (VISIGRAPP 2015). INSTICC, SciTePress. ISBN 978-989-758-088-8, pp. 29–39. DOI:10.5220/0005292900290039.

75. Dou W, Yu L, Wang X et al. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 2002–2011.
76. Luo D, Yang J, Krstajic M et al. Eventriver: Visually exploring text collections with temporal references. *IEEE Transactions on Visualization and Computer Graphics* 2012; 18(1): 93–105. DOI:10.1109/TVCG.2010.225.
77. Lee MR and Chen TT. Revealing research themes and trends in knowledge management: From 1995 to 2010. *Knowledge-Based Systems* 2012; 28: 47–58.
78. Alsakran J, Chen Y, Luo D et al. Real-time visualization of streaming text with a force-based dynamic system. *IEEE Computer Graphics and Applications* 2012; 32(1): 34–45. DOI:10.1109/MCG.2011.91.
79. Cui W, Liu S, Tan L et al. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics* 2011; 17(12): 2412–2421. DOI:10.1109/TVCG.2011.239.
80. Chavalarias D, Quentin L and Delanoë A. Replication Data for: Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies, 2021. DOI: 10.7910/DVN/SBH3EI. URL https://doi.org/10.7910/DVN/SBH3EI.
81. De Saussure F. *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag, 1916.
82. Maple C. Geometric design and space planning using the marching squares and marching cube algorithms. In *2003 International Conference on Geometric Modeling and Graphics, 2003. Proceedings*. IEEE, pp. 90–95.
83. Darwin C and Bynum WF. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. Penguin Harmondsworth, 2009.
84. Rivers WHR. The genealogical method of anthropological inquiry. *The Sociological Review* 1910; 3(1): 1–12.
85. Koutsofios E and North SC. Drawing graphs with dot 1996; .
86. Fry BJ. *Computational information design*. PhD Thesis, Massachusetts Institute of Technology, 2004.
87. Lobbé Q, Alexandre D and Chavalarias D. Replication data : Exploring, browsing and interacting with multi-scale structures of knowledge, 2021. DOI:10.7910/DVN/WLI9B5. URL https://doi.org/10.7910/DVN/WLI9B5.
88. Lobbé Q. Replication data : Exploring, browsing and interacting with multi-scale structures of knowledge, 2021. DOI:10.7910/DVN/4FQIA9. URL https://doi.org/10.7910/DVN/4FQIA9.
89. Terzopoulos D. Co-occurrence analysis of speech waveforms. *IEEE transactions on acoustics, speech, and signal processing* 1985; 33(1): 5–30.
90. Nguyen VT, Rivière P, Ripoll P et al. Research response to coronavirus disease 2019 needed better coordination and collaboration: a living mapping of registered trials. *Journal of Clinical Epidemiology* 2021; 130: 107–116. DOI:10.1016/j.jclinepi.2020.10.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0895435620311495.
91. Lobbé Q. Replication data : Exploring, browsing and interacting with multi-scale structures of knowledge, 2021. DOI:10.7910/DVN/SQULXL. URL https://doi.org/10.7910/DVN/SQULXL.
92. Tomita E, Tanaka A and Takahashi H. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science* 2006; 363(1): 28–42. DOI:10.1016/j.tcs.2006.06.015. URL https://www.sciencedirect.com/science/article/pii/S0304397506003586.
93. Dias G, Mukelov R and Cleuziou G. Mapping General-Specific Noun Relationships to WordNet Hypernym/Hyponym Relations. In *Knowledge Engineering: Practice and Patterns*. Springer, 2008. pp. 198–212. 00003 bibtex: dias2008mapping.
94. Blondel VD, Guillaume JL, Lambiotte R et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008; 2008(10): P10008. DOI:10.1088/1742-5468/2008/10/P10008. URL https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008. Publisher: IOP Publishing.

A The literature of text analysis and knowledge visualization

In order to extract the scientific literature of *text analysis and knowledge visualization* from the *Web of Science*'s online database, we have used the following queries:

"mind map" OR "topical map" OR "knowledge map" OR "science map" OR "science mapping" OR "mapping science" OR "mapping of science" OR "semantic map" OR "co-word" OR "co-citation" OR cocitation OR "co-term" OR "concept map" OR "information cartography" OR "mapping research" OR "visualization of knowledge" OR "bibliographic coupling" OR "citation analysis" OR "topic modeling" OR "latent dirichlet" OR (("text-mining" OR "text-analytics") AND (visualization OR infoviz OR "visual analytics")).

And retrieved the meta-data from a corpus of 13844 papers published between the '80s and April 2020. Please note that before 1990, most of the time, abstracts are missing in the meta-data.

Based on both automatic extraction and human pruning, the free software Gargantext (see 3.6) identifies a list of 876 terms constituting the specific vocabulary enclosed in our targeted corpus. The software then generates a semantic landscape appearing as a network of terms (i.e., nodes). In addition, the weighted edges represent any similarities found between terms; they are processed regarding the terms' co-occurrences in the corpus. We have elected the *confidence* similarity metric[93], a good proxy to measure direct interactions between elements of vocabulary. A map based on the confidence metric might indeed reveal the topology of the relationships between *hypernymies* (i.e., superordinate grouping terms) and *hyponymies* (i.e., terms more specific than others), both in terms of meanings and uses. Such a map is thus able to reconstruct a semantic area made of general concepts linked to more specific notions. In our case, a community detection algorithm also highlights the main research domains[94] as displayed by Figure 2.

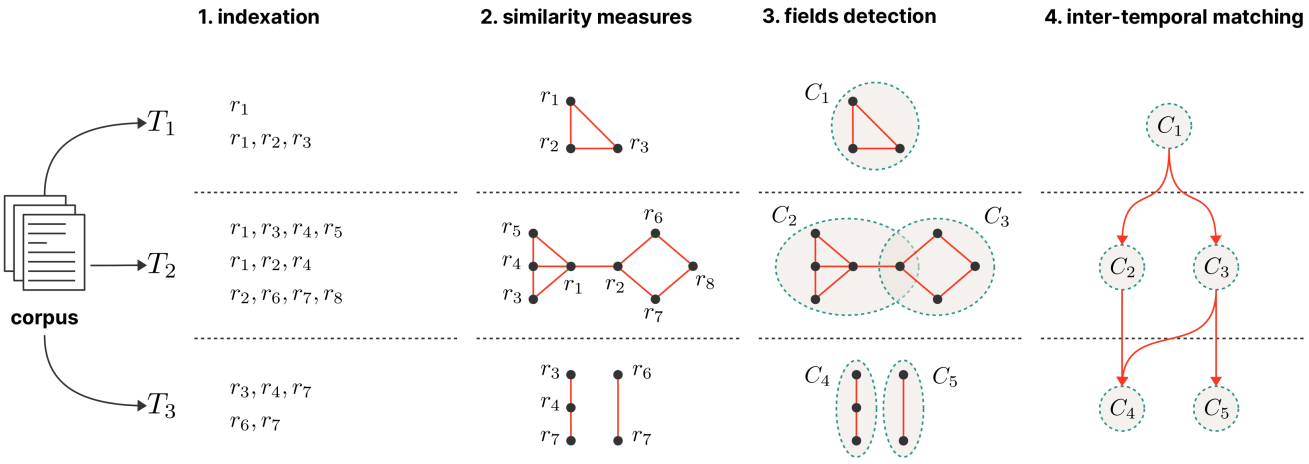

Figure 10. The four operators of reconstruction

B The phylomemy reconstruction process

The phylomemy reconstruction process[14] is part of the larger family of reconstruction methods (see 1) and is based on co-word analysis approach (see 2.1). Within the scope of the generic chain $O \rightarrow R \rightarrow V$, a phylomemy is a formal object R designed to reveal conceptual lineages out of any kind of unstructured but timestamped corpora of texts O. Here, the transition Φ from O to R can be described as a combination of four operators $\Phi = {}^4\Phi \circ {}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$[15], where:

${}^1\Phi$. **Indexation.** We start by framing a corpus of textual documents from a given database (see 4) before extracting its core vocabulary as a list $\mathcal{L} = \{r_i \,|i \in \mathcal{I}\}$ of root terms r_i. We then choose a temporal resolution (e.g., 3 years) that discretizes our corpus into an ordered set of periods $\mathcal{T}^* = \{T_i\}_{1 \leq i \leq K}, T_i \subset \mathcal{T}$. For each period, we finally compute the co-occurrence of terms in the documents.

${}^2\Phi$. **Similarity measures.** We build upon the matrix of co-occurrences to compute a similarity measure (e.g., the confidence measure, see 2.1) that we use as a proxy to estimate the semantic similarity between the terms.

${}^3\Phi$. **Fields detection and clustering.** Within each period, the completion of ${}^2\Phi \circ {}^1\Phi$ results in a graph of similarity potentially containing meaningful sub-units of terms C^T called *fields*. We then use of specific clustering algorithms (like *frequent item set* or *maximal clique*[15]) to identify these fields as sets of groups of terms over the periods, such as $\mathcal{C}^* = \{\mathcal{C}^T | T \in \mathcal{T}^*\}$ with $\mathcal{C}^T = \{C_j | j \in J^T\}$ and $C_j = \{r_i \,| r_i \in \mathcal{L}, i \in \mathcal{I}_j \ \subset \mathcal{I}\}$.

${}^4\Phi$. **Inter-temporal matching.** Finally, an inter-temporal matching mechanism reconstructs any kinship connections between groups from one period of time to another. It tries to assign each group a set of parents and children and, by doing so, highlights elements of conceptual and semantic continuity over time called *branches*. This resulting structure of terms, groups, links and branches determines the overall shape of the phylomemy.

Figure 10 summarizes linkages between ${}^1\Phi, {}^2\Phi, {}^3\Phi$ and ${}^4\Phi$. But the comprehension of the last mechanism requires a fine-grained explanation. Indeed, ${}^4\Phi$ first relies on an inter-temporal matching function that derives from a similarity measure $\Delta : \mathcal{C} \times \ \mathcal{P}(\mathcal{C}) \rightarrow [0, 1]$ (such as a *Jaccard coefficient*[93]). This function aims to create upstream/downstream kinship connections between a given group C^T at period T and any single/pairs of candidates $\{C_j\}_j \subset \mathcal{P}(\mathcal{C})$ belonging to a strictly anterior/superior period T'. The resulting lineages are then validated if and only if their corresponding values of Δ satisfy a fixed threshold $\delta \geq 0$.

C Data sets

All the corpora and lists of terms used in this paper have been described in[15] and are available in[80]. The reconstructed phylomemies can be downloaded as archives: $\mathcal{D}_{cnrs}$ in[87]; $\mathcal{D}_{maps}$ in[88]; $\mathcal{D}_{ct}$ in[91].