Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies

David Chavalarias, Quentin Lobbe, Alexandre Delanoë

Highlights

Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies

David Chavalarias[1,2,*],Quentin Lobbé[2],Alexandre Delanoë[2]

- we formalize the notion of *level* and *scale* of knowledge dynamics as complex systems,
- we propose a new class of meaning for the reconstruction of knowledge dynamics formalized by a new objective function parameterized by the level of observation,
- we properly formalize the concept of phylomemy as distinct from the concept of phylomemetic networks,
- we propose a new reconstruction algorithm for phylomemetic networks reconstruction that outperforms previous ones thanks to a new objective function,
- we show in case studies that this approach produces representations of knowledge dynamics close to the ones that can be obtained by synthesizing the points of view of experts on a given domain (glyphosate literature and knowledge and science mapping literature),
- we demonstrate with cases studies that this approach can be applied to any kind of unstructured corpora, even on relatively small data sets or short texts,
- we propose a new temporal clustering on dynamical graphs that is naturally part of the process of multi-level and multi-scale reconstruction of phylomemies,
- we integrate user preferences into our framework by providing an interaction model and contextualizing the different elements of our reconstruction workflow in the theoretical framework of the embodied cognition,
- By applying our method to the state-of-the-art of this paper, we illustrate how it could be systematically applied to generate such extended historical analysis which might be helpful to give more context to the scope and contributions of scientific papers.

Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies

David Chavalarias[1,2,*], Quentin Lobbé[2] and Alexandre Delanoë[2]

(1) Centre d'Analyse et de Mathématiques Sociales, CNRS/EHESS, (2) Complex Systems Institute of Paris Île-de-France (ISC-PIF), CNRS, Paris, France. (*) Corresponding author.

ARTICLE INFO

Keywords:
phylomemy reconstruction
knowledge dynamics
phenomenological reconstruction
multi-scale and multi-level complex systems
science map
co-word analysis

Abstract

The little prince asked Saint-Exupéry to draw him a sheep, but what if he had asked him to be drawn Science? How could he have done it and what could we have learned from it? In this article, we address the question of "drawing science" by taking advantage of the massive digitization of scientific production, and focusing on its body of knowledge. We demonstrate how we can reconstruct, from the massive digital traces of science, a reasonably precise and concise approximation of its dynamical structures that can be grasped by the human mind and explored interactively. For this purpose, we formalize the notion of *level* and *scale* of knowledge dynamics as complex systems and we introduce a new formal definition for phylomemetic networks as dynamical reconstruction of knowledge dynamics. We propose a new reconstruction algorithm for phylomemetic networks that outperforms previous ones and demonstrate how this approach also makes it possible to define a new temporal clustering on dynamical graphs. Finally, we show in case studies that this approach produces representations of knowledge dynamics close to the ones that can be obtained by synthesizing the points of view of experts on a given domain.

1 Introduction

1.1 The shapes of science

The little prince asked Saint-Exupéry to draw him a sheep, but what if he had asked him to be drawn Science? How could he have done it and what could we have learned from it?

Producing a global representation and understanding of the knowledge mankind has produced so far has been multi-century-old quest. In 1751, Jean le Rond d'Alembert, in his introduction of the first French Encyclopédie, stated his ambitions "to make a genealogical or encyclopedic tree which will gather the various branches of knowledge together under a single point of view and will serve to indicate their origin and their relationships to one another"[1]. Since then, several disciplines have investigated this question, from history and philosophy of science [18] to the emerging field of science mapping [21].

Science can be defined quite generically as "1) body of knowledge, 2) method, and 3) way of knowing" [1]. This body of knowledge is the result of the distributed interactions of thousands of scientists over the years; currently producing about two million documents per year - journal articles alone.

In this article, we will address the question of "drawing science" by taking advantage of the massive digitization of scientific production, and focusing on its body of knowledge. We demonstrate how we can reconstruct, from the massive digital traces of science, a reasonably precise and concise approximation of its structures that can be grasped by the human mind and explored interactively (an operation called *phenomenological reconstruction*[2] [5, 18]). Science being a decentralized and dynamical process [37], complex systems methods are increasingly involved in the research

[1] "former un Arbre généalogique ou encyclopédique qui les rassemble sous un même point de vûe, & qui serve à marquer leur origine & les liaisons qu'elles ont entre elles."

[2] Phenomenological reconstruction is the process of choosing the appropriate data to be collected about phenomena and pre-structuring them to allow for a more comprehensive understanding in subsequent analyses. Ideally, phenomenological reconstruction may provide us with candidate concepts and relations, which, when integrated into modeling, can then serve as a basis for the human experimental work.

domain of science of science (SoS) [75]. In this paper, we will model the semantic structures of science using *phylomemy reconstruction* a method proposed by Chavalarias and Cointet [16] we are now extending to make it both more accurate and naturally multi-level and multi-scale. We will then demonstrate with some case studies how this achieves d'Alembert's dream: of "[...] collecting knowledge into the smallest area possible and of placing the philosopher at a vantage point, so to speak, high above this vast labyrinth, whence 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."[32].

Finally, since the proposed methodology requires only time-stamped textual content, it can be applied to the reconstruction of knowledge dynamics from any kind of corpora, which makes this approach very general and able to provide knowledge on any human activity, provided it leaves a digitalized written trace (e-mails, reports, patents, news, webpages, microblogs, etc.).

1.2 Levels and scales in complex systems

Complex systems display structures at all scales with a hierarchical organization reflecting the interactions between the entangled processes that sustain them [14]. Their description mobilizes the notions of 'levels' and 'scales', "level being generally defined as a domain higher than 'scale"' and 'scale' referring to the structural organization within a level [42]. In biology for example, the choice of level of observation determines what the main entities under study (individuals, organs, cells, genes, etc.) are, while the choice of a scale determines the smallest resolution adopted to describe these entities.

The method presented in this paper makes a clear distinction between these notions of *level* and *scale* in the phenomenological reconstruction of knowledge dynamics[3], a distinction that is not made explicit by other scholars or science modeling approaches. The choice of a *level* of observation determines the range of intrinsic complexity of the dynamic entities we want to observe, the choice of a *scale* defines the extrinsic complexity of their description[4]. One of the main difference between *level* and *scale* is that the concept of level is ontologically linked to the notion of time since the components of a level derive their unity from some underlying dynamic process ; while the notion of scale does not necessarily imply time.

1.3 Levels of observation of Science

At a given level of observation, science is composed of different research domains whose unity over time is characterized by core questions and research objects. For example, at a very high level, one may find *social sciences*, *biology*, *physics*, etc.; at a lower level, one will find *sociology*, *cell biology*, *quantum mechanics* ; and at an even lower level, *political sociology*, *proteomics* or *quantum cryptography*.

Levels are sustained by socio-economic processes that guide the progress of science with more or less of a broad focus (laboratories, national and international research networks, national and international funding schemes, journals, etc.). Their sub-components are evolving socio-semantic macro-structures that are reflected by the specialization of certain journals, conferences or the presence of sub-categories in digital archives.

In this article, we will propose a definition of what the components of a level of organization of Science are. We will call these components *branch of science*. We will also detail a method to identify and represent them.

2 Related work

2.1 Mapping science and knowledge

Producing *science maps* and *knowledge maps*, as our in-depth analysis of 14k papers[5] will show (see SI Appendix B and SI Figures 13 and 12), has been addressed by two scientific approaches, each of them having known recent developments regarding evolution, dynamics and temporality.

[3] Note that some scholars use the term 'resolution', 'zoom' or 'granularity' instead of 'scale'.

[4] Are we only interested in the main concepts of the fields or in a finer granularity of a myriad of terms? Do the details of the interactions between terms matter? Does the scientific field under study evolve linearly or with ramifications? etc.

[5] 'Co-word analysis of 14,374 documents retrieved on 2020 04 27 from the Web of Science with the query : "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 "mapping research" OR "visualization of knowledge" 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 analyzed with the Gargantext (http://gargantext.org) free software.

1. **Bibliometrics** is defined as the use of statistical methods to analyze digitized textual corpora. The part of this domain dealing with science mapping is divided into two major components:

   - *Citation and co-citation analysis (CA)* emerged in the 1970s [34] with the aim of measuring and analyzing scholarly literature. Primarily focused on the assessment of scientific output, it quickly diversified with the analysis of large citation landscapes through methods such as *co-citation* and *bibliographic coupling* [40, 57]. Later on, following the creation of the *Web*, these methods were generalized as part of the study of *hyperlinked data* [41]. Methods to describe *conceptual structures* of science such as *research fronts*, *hot topics* and *trends*, etc. [72, 10] came at the forefront of this research domain. Over the last decade, a growing number of contributions have proposed temporal reconstructions of the citation landscape [**?** 24, 13].
   - *Co-word analysis (CWA)* is a bottom-up approach first developed by sociologists in the 1980s [12] to reconstruct the dynamics of *research themes* out of words *co-occurrence*. It has developed in the last decade into a generic approach to map knowledge dynamics in unstructured corpora [16, 69, 53].

   Citation and co-word analysis research areas share common objectives, namely to understand the structures of science. This is reflected by their proximity on the map of Figure 12. They form a toolkit that is at the core of the bibliometrics approach to *science of science*. Their parallel development has quickly paved the way to hybrid research between co-word and citation analysis [7, 6] or social and semantic networks [52]. Nowadays, science maps and knowledge maps are interdisciplinary objects of research resulting from both quali-quantitative and socio-technical processes [22]. Finally, the growth of scientific databases has stimulated the visualization of wide *citation landscapes* [58] or complex atlases of sciences [8, 9].

2. **Information retrieval (IR)** deals with the retrieval and evaluation of information from digitized document repositories. It has mobilized latent semantic analysis and topic modeling in the early 2000s at the instigation of a community of statisticians who introduced the *latent dirichlet allocation* method [3] to primarily characterize *collections of documents*. Even though they mostly focus on *document classification* [71], *recommendation* [68] or *sentiment analysis* [44], they have started in some of their most recent work to tackle the mapping of scientific issues and their dynamics [47, 74, 39].

More recently, the idea of working on latent semantic spaces has been taken up by the research field of machine learning with approaches such as *words embedding* [64, 62]. It been mostly used for the purpose of information retrieval and document classification (which is the reason why it appears in the same group as LDA in the SI map Figure 13), but can also be a useful tool to analyze science's evolution, mostly at the micro-level of terms-to-terms similarities [50, 64, 62].

2.2 Mapping knowledge dynamics

Work in the field of information retrieval (IR) has primarily sought to respond to queries on large textual databases. The subdomain of IR related to our work mainly aims to visualize the evolution of topics related to a specific query and to proceed to a different reconstruction for each query. Moreover, this previous research has been focusing specifically on the document level, the objective being to classify a set of retrieved documents according to their topics and temporal relationships in order to produce streams of related papers [43, 54]. Nevertheless, some recent works in this area do propose the reconstruction of knowledge dynamics at the scale of a corpus [27, 36, 74, 55].

As discussed in section 3.2.3, our methodology belongs to the bibliometric branch of science of science, whose first objective is the reconstruction of scientific landscapes and their dynamics. In order to highlight the advantages and disadvantages of the methods issued from these various approaches, we have identified some of their main characteristics in SI section B.2.

We have used these characteristics in Table 1 and Table 2, to compare some key papers from each methodological approach. Except for the method proposed in this paper, few previous studies meet all of them. In particular, none of them really formalize the distinction between the notions of 'level' and 'scale'.

Table 1
Comparison between different approaches for the modeling of knowledge dynamics. The column 'meaning' indicates whether the approach can distinguish between paradigmatic (P) and syntagmatic (M) relations among terms, process both or one in particular, or cannot make this distinction at all. Symbol × means that the feature is not part of the study or not compatible with the approach. *Abbreviations. Domains of origin:* Science of Science (SoS), Information Retrieval (IR), Machine Learning (ML), Visual Analytics (VA) ; *Method:* Complex Network Analysis (CNA), Co-Word Analysis (CWA), Word Embedding (WE), Topic Modeling (TM), Citation Analysis (CA)

Paper	Domain	Focus	Unit of analysis	Methods	Fields detection	Meaning
This paper	SoS	corpora based	text	CWA	COC	P+S
[53]	SoS	corpora based	text	CWA	COC	P+S
[69]	SoS	corpora based	text	CWA	COC	P+S
[16]	SoS	corpora based	text	CWA	COC	P+S
[55]	IR	corpora based	text	CWA	COC	P+S
[43]	IR	corpora based	documents	CNA	MGF	×
[62]	ML	corpora based	codes	WE	×	P
[50]	ML	corpora based	codes	WE	×	P
[39]	IR	corpora based	documents	TM	DTM (LDA)	P
[20]	IR	corpora based	documents	TM	LDA	P
[74]	Visual analytics	corpora based	documents	TM	HLMT	×
[67]	IR	corpora based	documents	TM	dDTM	-
[27]	IR	corpora based	documents	TM	HDP	×
[36]	IR/ML	corpora based	documents	TM	PLSA	×
[13]	SoS	corpora based	documents	CA	BC	×
[24]	SoS	corpora based	documents	CA	BC	S
[38]	SoS	query based	documents	TM + CA	PLS + CA	×
[54]	IR	query based	documents	IR	×	×

Table 2
Comparison between different approaches for the modeling of knowledge dynamics. Legend. ✓: the property is fully part of the study, ⋯: the property is not part of the study but further developments could in principle integrate this aspect, ×: the feature is not part of the study or not compatible with the approach, ?: the elements given in the paper did not allow us to assess this aspect, -: this criteria is irrelevant for this paper. *Method:* Complex Network Analysis (CNA), Co-Word Analysis (CWA), Word Embedding (WE), Topic Modelling (TM), Citation Analysis (CA)

Paper	This paper	[53]	[69]	[16]	[55]	[43]	[62]	[50]	[39]	[20]	[74]	[67]	[27]	[36]	[13]	[24]	[38]	[54]
Publication year	2021	2015	2014	2013	2013	2019	2020	2019	2018	2017	2017	2015	2011	2009	2020	2016	2011	2012
Method	CWA	CWA	CWA	CWA	CWA	CNA	WE	WE	TM	TM	TM	TM	TM	TM	CA	CA	TM+CA	IR
Advanced Text-mining ?	✓	✓	×	✓	×	×	×	×	×	×	?	×	×	y	×	×	×	×
Work on unstructured text ?	✓	✓	✓	✓	✓	×	⋯	⋯	✓	✓	✓	✓	✓	×	×	×	×	✓
Work on short texts ?	✓	✓	✓	✓	✓	⋯	✓	✓	?	×	?	×	×	×	✓	n/a	?	×
Work on less than 10k doc ?	✓	✓	✓	✓	✓	✓	×	×	?	?	?	?	?	?	✓	✓	?	✓
Global science maps ?	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×
Unconstrained # of topics	✓	✓	✓	✓	✓	✓	-	-	×	×	✓	×	✓	×	✓	✓	✓	-
Evolving topics ?	✓	✓	✓	✓	✓	✓	-	-	✓	✓	×	✓	✓	✓	✓	✓	✓	-
Allow re-emerging topics ?	✓	×	×	✓	✓	✓	-	-	×	×	✓	×	×	×	×	×	✓	-
Allow split/merge events	✓	✓	✓	✓	✓	✓	-	-	×	✓	×	×	✓	×	✓	✓	✓	✓
Multi-level	✓	×	×	⋯	⋯	×	×	-	×	×	×	×	×	×	×	×	⋯	⋯
Multi-scale	✓	⋯	✓	✓	✓	✓	×	-	✓	×	✓	✓	✓	×	✓	×	×	✓
Objective function	✓	×	×	×	✓	✓	-	-	✓	×	✓	✓	×	✓	✓	×	×	✓
External validation	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	✓	×	✓	×	×	✓
Internal / quantit. evaluation	✓	×	×	✓	✓	✓	✓	✓	✓	×	✓	✓	×	×	✓	×	×	✓
Integrate users preferences ?	✓	×	/	×	⋯	⋯	×	×	×	×	×	×	⋯	×	×	×	×	✓
Advanced visualizations	✓	✓	✓	✓	✓	×	×	×	×	✓	✓	×	✓	✓	✓	✓	✓	✓
Software reproducibility	✓	✓	?	?	?	×	×	×	?	×	?	×	?	×	✓	×	×	?
Open Source	✓	?	?	×	?	×	×	×	?	×	?	×	?	×	✓	×	×	?

2.3 Phenomenological reconstruction of science with phylomemies

To observe and further understand through modeling a complex object $O \in \mathcal{O}$, we first select the properties to be observed and measured, then reconstruct from the data collected those properties and their relations as a formal object $R \in \mathcal{R}$ described in a high-dimensional space. Then some dimension reduction is applied to R to get a human-readable representation in a space $\mathcal{V}$. The chain $\mathcal{O} \longmapsto \mathcal{R} \longmapsto \mathcal{V}$ defines what is called *phenomenological reconstruction* [5, 18]. The quality of a phenomenological reconstruction is measured by its ability to propose, from the raw data, representations in $\mathcal{V}$ that make sense to *us* and provide affordances for modeling and conceptual understanding[6].

Phylomemy reconstruction is a general methodology for the processing of unstructured and dynamical textual corpora that is part of the co-word analysis framework[11]. It can be applied to any kind of corpora, from scientific papers [16, 17] to news papers [19] or political discourses [53], in order to identify the topics they cover and the dynamical structures generated by the development of these topics.

In this paper, we will define formally a category of meanings conveyed by *phylomemies* [16] as phenomenological reconstructions. We will then show how this definition can be made operational and allow users to naturally grasp the multi-level and multi-scale structure of knowledge dynamics through the definition of 1) $\mathcal{O} \longmapsto \mathcal{R}$, the operator that reconstructs a multi-level and multi-scale structure, and 2) $\mathcal{R} \longmapsto \mathcal{V}$, projector that selects a level of observation and propose to explore this level in a multi-scale way. At the same time, we makes tangible the different shapes generated by science, which can be visualized with an appropriate free software (cf. [45]). More discussions on the relations between the phenomenological reconstruction of science with phylomemies, formal modeling and history and philosophy of science can be found in [18].

3 Materials and methods

3.1 Generic workflow of phylomemy reconstruction

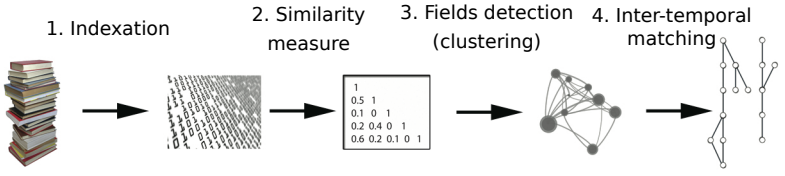

Figure 1: Workflow of phylomemy reconstruction from raw data (digitized textual corpora) to global patterns. The output is a set of phylomemetic branches where each node is constituted by a network of terms describing a research field. These nodes are a proxy of scientific fields and can have different statuses: emergent, branching, merging, declining. Source: [16]

The workflow of phylomemy reconstruction (cf. Figure 1 and [16]) takes as input a large set of documents $\mathcal{D}$ produced over a period of time $\mathcal{T}$ (the raw data in $\mathcal{O}$), and provides as output a structure that characterizes, at a given spatio-temporal resolution, the transformations of the knowledge domains covered by $\mathcal{D}$. An example of such output can be seen in the phylomemy of glyphosate-related academic literature Figure 2 and is detailed as a case study in SI section C.

The workflow for phylomemy reconstruction uses advanced text-mining and complex networks analysis. It can be roughly decomposed into four operators $\Phi = {}^4\Phi \circ {}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$ that correspond to four main steps (see Figure 1):

${}^1\Phi$. **Indexation.** The core vocabulary of $\mathcal{D}$ is extracted as a list $\mathcal{L} = \{r_i \mid i \in \mathcal{I}\}$, where r_i are groups of terms (thereafter called *roots*) conveying the same meaning according to the analyst[7]. An ordered series $\mathcal{T}^* = \{T_i\}_{1 \leq i \leq K}$, $T_i \subset \mathcal{T}$, of sub-periods of $\mathcal{T}$ is defined to determine the temporal resolution of the reconstruction. Co-occurrences of *roots* within the documents are then processed per period of time.

${}^2\Phi$. **Similarity measures.** Each period-dependent, root-to-root co-occurrence matrix is transformed into a root-to-root similarity matrix after the appropriate similarity measure has been chosen. This choice should be oriented

[6] Even though hypotheses about the underlying processes that have generated the data could help to find the relevant phenomenological reconstruction method, a phenomenological reconstruction does not make such hypotheses.

[7] These groups could be obvious e.g. *{decision-making processes, decision making process, decision making processes}* or more customized to the analyst's point of view, e.g. *{advertising, advertisement, advertisements, advertiser}*.

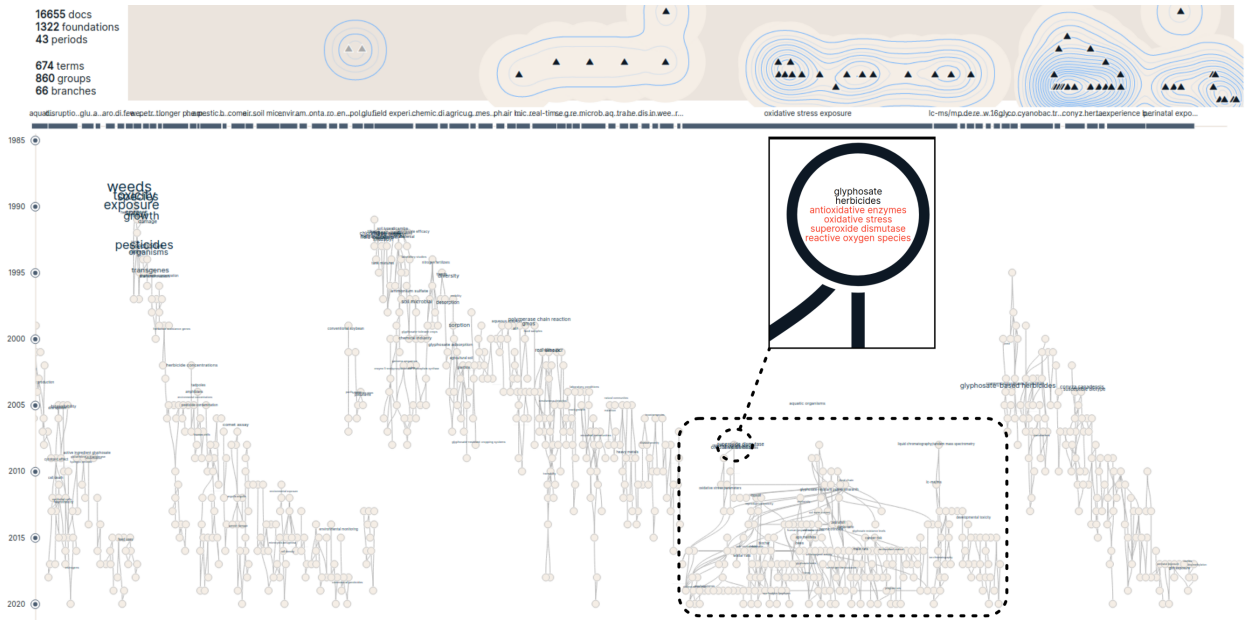

Figure 2: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ (16,655 documents) at the level $\lambda = 0.8$ between 1995 and 2020 with branches smaller than 3 filtered out. Each connected dot represents a specific field of research described by several key-words (as displayed for example in the magnified area: {*glyphosate, herbicides, antioxidative enzyme, oxidative stress, superoxide dismutase, reactive oxygen species*}). Emerging terms are highlighted in red in the description of the field and are displayed in large fonts on this figure, close to the earliest fields that mention them. Inter-temporal matching between fields is represented by vertical links and a group of connected fields defines a branch of science. The branch highlighted by a dotted box starts in 2006 and deals with majors negative side-effects of glyphosate-based products. It was indeed in the 2000s that significant discrepancies were documented regarding the decomposition and residues of glyphosate, between laboratory experiments on glyphosate and field experiments on glyphosate-based products, raising concerns about their effects on health and on the environment. At this level of observation, glyphosate research displays 66 branches composed of a total of 860 fields and covering 674 roots of the original list $\mathcal{L}$ of 1322 roots. An interactive version of this phylomemetic network is available at `http://unpublished.iscpif.org/glyphosate`. Details about this case study are given in SI section C and an interactive version of this visualization can be downloaded from the archives [60] (data), [59] (explorer). Full details about this method to explore phylomemetic networks are available in [45].

by the research question at stake, as described for example in [29] or [70]. Depending on the question, two reconstructions using two distinct classes of similarity measures may prove complementary (see also SI. B).

$^3\Phi$. **Fields detection and clustering.** Within each period, the completion of $^2\Phi \circ {}^1\Phi$ results in temporal series of networks of roots similarities. A community detection algorithm[8] is then applied to identify, within these networks and for each period of time $T_i \subset \mathcal{T}$, important sub-units constituting "fields of knowledge", *i.e.* dense networks of 'key-words' characterizing key research questions. Research on community detection algorithms has been very prolific [33] and several algorithms with different intrinsic spatial resolution could potentially be suitable at this step. The result of fields detection is a set of clusters of roots C_j computed over $\mathcal{T}^*$: $C^* = \{C^T | T \in \mathcal{T}^*\}$ where $C^T = \{C_j | j \in J^T, \}$ and $C_j = \{r_i \,|r_i \in \mathcal{L}, i \in \mathcal{I}_j \ \subset \mathcal{I}\}$. We will note $\mathcal{C} = \bigcup_{C^T \in C^*} C^T$ the set of all clusters over all periods.

$^4\Phi$. **Inter-temporal matching.** Once the fields of knowledge have been identified at a given spatio-temporal resolution, inter-temporal matching reconstructs the lineage between these fields. The most general assumption on the evolution of fields of knowledge is that some are likely to to emerge, split, merge or die ([49] for pioneer work or [51] for a review about community dynamics in evolving networks). The problem can be formulated as follows: find, for any cluster C^T at period T, the combination of clusters from previous periods, if any, that could best

[8] Also called *graph clustering* algorithm.

account for the presence of C^T at T (the 'parents' of C^T), as well as the set of clusters from subsequent periods that could be the continuation of cluster C^T as a knowledge field (the children of C^T). The result of this step is an evolving structure describing the evolution of large knowledge domains and it has been called a *phylomemetic network*[16]. Inter-temporal matching of a temporal series of clusters brings time into play. Consequently, this is where the notion of *level* arises.

Steps 1 to 3 are very common steps in science and knowledge mapping literature in which, at each stage, several options are available to the modeler according to his/her initial research questions[9]. In the following sections, we will focus on ${}^4\Phi$ and how it can convey the notions of *level* and *scale*, setting arbitrary parameters for ${}^1\Phi$ to ${}^3\Phi$. The reader interested in these particular steps can find more examples and technical details in [16, 25, 15, 28] for more examples and technical details.

3.2 Method of reconstruction of the dynamics

Processing the weighted inter-temporal links between elements of a temporal series of clustering is the easy part. Once 'thresholded', temporal networks always can be interpreted as more or less reflecting dynamic macro-structures. There is however some subtlety here.

First, since the temporal threshold has a direct impact on the overall connectivity of the dynamic structure, and consequently on the number of sub-components, the interpretation of the final result depends strongly on the procedure for setting the threshold.

Second, since the goal is to highlight the continuities in the evolution of clusters, parents and children of a particular cluster will be looked for as close in time as possible. Consequently, when allowing matching between non-consecutive periods, the thresholding operation also influences the average time difference between two related fields in the reconstruction.

As we will see, this entanglement between the granularity of the macro-structures revealed by a temporal reconstruction and the timescale of inter-temporal matching is where the multi-level aspect of phylomemy comes into play.

3.2.1 Upstream and downstream inter-temporal matching

An up-stream inter-temporal matching function is based on a similarity measure $\Delta : \mathcal{C} \times \mathcal{P}(\mathcal{C}) \rightarrow [0, 1]$ that defines the 'strength of association' between any clusters $C^T \in \mathcal{C}$ and any set of clusters $\{C_j\}_j \subset \mathcal{P}(\mathcal{C})$ belonging to strict anterior periods T' (noted $T' \prec\prec T$) [10]. Starting from a cluster similarity function Δ, Chavalarias and Cointet [16] proposed to find for every period $T \in \mathcal{T}^*$, for every cluster C^T computed over the period T and for every threshold $\delta \geq 0$ the closest satisfactory set of 'parents' ${}^4\Phi^{\prec}_{\delta}(C^T)$ according to Δ with their association strength:

$$ {}^4\Phi^{\prec}_{\delta}(C^T) = (\{C_j \in \kappa^{\prec}_{C^T}\}, w) $$

Where:

- $\kappa^{\prec}_{C^T} = \arg max_{\Delta(C^T,\kappa)}[\arg min_{\{\tau(C^T,\kappa)|\kappa \subset \mathcal{C}^{T' \prec\prec T}, \Delta\left(C^T,\kappa\right) \geq \delta\}} \kappa]$,
- $\mathcal{C}^{T' \prec\prec T} = \{C^{T'} \in \mathcal{C} | T' \prec\prec T\}$ is the set of all clusters of $\mathcal{C}$ whose period is strictly anterior to T,
- $\tau(C^T, \kappa)$ is the minimum amount of time elapsed between the period T and the periods of the clusters constituting κ[11].
- $w = \Delta\left(C^T, \kappa^{\prec}_{C^T}\right) \in [0, 1]$ is the association strength between C^T and its "parents".

[9] For example, step 1.2 can proceed from advanced text-mining on the corpora, or be made via external ontologies or thesaurus (*e.g.* PubMed Mesh).

[10] $\mathcal{P}(X)$ is the set of all parts of X.

[11] Their approach makes it possible to have several sets of parents for a given cluster although this situation is quite rare. Also, since a given cluster could be in the set of parents of several subsequent clusters, a cluster might have many children.

As the definitions of parents and children of a cluster are symmetrical with respect to time, inter-temporal matching is consequently symmetrized by considering both the up-stream ${}^4\Phi^{\prec}_{\delta}(C^T)$ and the symmetric down-stream inter-temporal matching function ${}^4\Phi^{\succ}_{\delta}(C^T)$, where the association strength between any clusters $C^T \in \mathcal{C}$ and any set of clusters $\{C_j\}_j \subset \mathcal{P}(\mathcal{C})$ belonging to strict posterior period T'' (noted $T'' \succ\succ T$) is also processed (see SI D.4 for full description).

We thus define ${}^4\Phi : \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(\mathcal{C}), w)^2$ defined as: ${}^4\Phi_{\delta}(C) = (({}^4\Phi^{\prec}_{\delta}(C), {}^4\Phi^{\succ}_{\delta}(C))$. It is worth mentioning that this function looks for the first correspondence that, from a temporal point of view, satisfies a given threshold δ, instead of looking for all potential correspondences for all time and taking the optimum - an approach adopted among all the works referenced in Table 2 that are going beyond the simple correspondence between consecutive periods.

3.2.2 *Phylomemies as foliation on temporal series of clustering*

Thereafter, we will work with the following definitions:

Definition: ***Temporal series of clustering*** $\mathcal{C}^*$**.** Let $\mathcal{T}^*$ be an ordered set and $\mathcal{L} = \{r_i \,|i \in \mathcal{I}\}$ be a set of elements. A temporal series of clustering over $\mathcal{L}$ is defined as $\mathcal{C}^* = \{C^T | T \in \mathcal{T}^*\}$ where $\mathcal{C}^T = \{C^T_j | C^T_j \subset \mathcal{P}(\mathcal{L})\}_{j \in J^T}$.

Definition: ***Foliation on a temporal series of clustering*** (cf. Figure 3). Let $\mathcal{C}^* = \{C^T | T \in \mathcal{T}^*\}$ be a temporal series of clustering and $\mathcal{C} = \bigcup_{C^T \in \mathcal{C}^*} C^T$, a foliation on $\mathcal{C}^*$ is defined as a function $\Phi :: \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(\mathcal{C}) \times [0,1])^2$ such as:

1. $\forall C \in \mathcal{C}^T, \forall \delta \in [0,1], \Phi(C,\delta)(1,1) \subset \mathcal{P}(\mathcal{C}^{T' \prec\prec T})$ (parents of $\mathcal{C}^T$ at δ, associated with strength $\Phi(C,\delta)(1,2)$),
2. $\forall C \in \mathcal{C}^T, \forall \delta \in [0,1], \Phi(C,\delta)(2,1) \subset \mathcal{P}(\mathcal{C}^{T' \succ\succ T})$ (children of $\mathcal{C}^T$ at δ, associated with strength $\Phi(C,\delta)(2,2)$),

Definition: ***Phylomemy***. A *phylomemy* ϕ is a foliation on a temporal series of clustering $\mathcal{C}^*$. It describes, for any cluster C^T_j in temporal components of $\mathcal{C}^*$ and *any* threshold δ, the relevant inheritance linkages of C^T_j. Consequently, for the study of knowledge dynamics, $\mathcal{R}$ is the space of all foliations on temporal series of *roots* clustering.

Definition: ***Weighted inheritance networks*****.** Let $\mathcal{C}^* = \{\mathcal{C}^T | T \in \mathcal{T}^*\}$ be a temporal series of clustering. A weighted inheritance network $\varphi : \mathcal{C} \longmapsto (\mathcal{P}(\mathcal{C}) \times [0,1])^2$ is a function defined over the set of nodes $\mathcal{C} = \bigcup_{C^T \in \mathcal{C}^*} \mathcal{C}^T$ such as $\forall C^T_j \in \mathcal{C}; \varphi(C^T_j)(1,1) \subset \mathcal{P}(\mathcal{C}^{T' \prec\prec T})$ and $\varphi(C^T_j)(2,1) \subset \mathcal{P}(\mathcal{C}^{T' \succ\succ T})$[12].

Definition: ***Phylomemetic network***. Let ϕ be a phylomemy over $\mathcal{C}^*$ and $\Pi : \mathcal{C} \longmapsto [0,1]$, a *phylomemetic network* is defined by $\varphi_\Pi = \{(C^T_j, {}^4\Phi^{\prec}_{\Pi(C^T_j)}(C^T_j)) | C^T_j \in \mathcal{C}\}$. It is a *plaque* of the phylomemy ϕ that defines a weighted inheritance networks over a temporal series of root clusters. It is thereafter possible to visualize the inheritance patterns of φ_Π in $\mathcal{V}$ in order to understand the structure of ϕ (cf. Figure 2 and [45]).

$\mathcal{V}$ is consequently the space of all weighted inheritance networks of root clusters, and phylomemetic networks are elements of $\mathcal{V}$ that can be defined as a plaque of a particular phylomemy.

3.2.3 *Quality assessment of phylomemetic networks and levels of observation*

To assess the quality of a phenomenological reconstruction, [23] make the distinction between "*interpretation* as the facility with which an analyst makes inferences about the underlying data and *trust* as the actual and perceived accuracy of an analyst's inferences."

Reconstructions of science dynamics, regardless of the methodology (cf. section 2), generally offer rich opportunities for interpretation because one naturally interprets salient events in the reconstruction (*e.g.* split, merge, emergence and disappearance of fields/clusters) as related to the fate of scientific fields. This richness of interpretability is however rarely backed with a high level of *trust*. The distinction between *phylomemies* belonging to $\mathcal{R}$ and *phylomemetic lattices* that belongs to $\mathcal{V}$ gives more theoretical grounding to the assessment of *trust* of phenomenological reconstruction.

[12] We note $\varphi(C^T)(1,1)$ the first component of the first tuple of $\varphi(C^T) = ((\mathcal{X}, w_1), (\mathcal{Y}, w_2))$, *i.e.* $\mathcal{X}$, and $\varphi(C^T)(2,1)$ the first component of the second tuple, *i.e.* $\mathcal{Y}$.

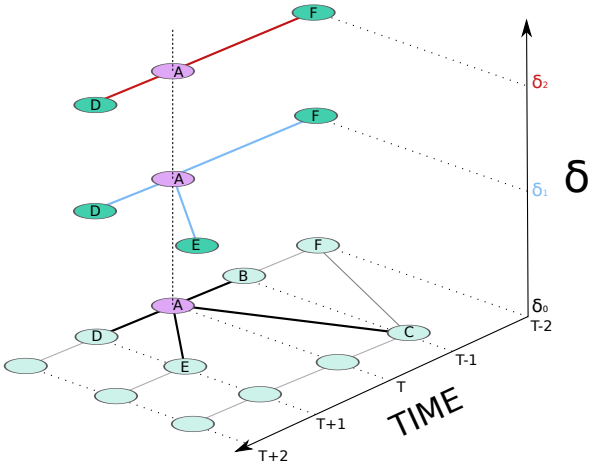

Figure 3: Local structure of a foliation on a temporal series of clustering $\mathcal{C}^*$ parameterized by δ. Locally, above the cluster $A \in C^T \subset \mathcal{C}^*$ from period T, δ parameterized the different sets of parents and children of A for different satisfaction threshold in inter-temporal matching. As we raise δ, the parents and children of A might change. At δ_0, clusters B and $C \in \mathcal{C}^{T-1}$ are the parents of A ; but when we raise δ to δ_1, the pair $\{B, C\}$ of parents is no longer valid for A and the cluster F from a former period $T-2$ has to be mobilized to describe the lineage of A as the child of $\{\mathrm{F}\}$. Eventually, some branches might split or merge due to these reconfigurations (*e.g.* here the segment starting from E becomes the starting of a new branch at δ_2). A phylomemy is a foliation over a temporal series of clustering $\mathcal{C}^*$.

In the framework of this generalization, [16] appear to have studied the specific case where the operators $\mathcal{R} \longmapsto \mathcal{V}$ are all uniform projectors $\Pi = \delta$, *i.e.* $\varphi_\delta = \{(C_j^T, {}^4\Phi_\delta^{\prec}(C_j^T)) | C_j^T \in \mathcal{C}\}$. The results obtained so far by [16] show that, for a wide range of δ values, there are robust correlations between the structural properties of fields assessed at specific period T, and their dynamical properties inferred from their position in the branches of φ_δ (*i.e.* emergent, declining, splitting, merging, etc. see Figure 2). This observation means that we can *trust* the interpretation our common sense associates to the fact that a field occupies some salient position in a phylomemy.

There is no reason to think however that a uniform projector Π_δ will provide the same level of description for all knowledge domains. Moreover, since no objective function has been proposed to help the analyst interpret and choose the value of δ, there are some concerns about the actual as opposed to perceived accuracy of inferences one can make when comparing branches of a phylomemy or interpreting their distribution.

To get around this problem, we propose a class of meanings for the operation $\mathcal{R} \longmapsto \mathcal{V}$ that is commonly understandable. In order to do so, let's come back to D'Alembert's dream to "distinguish the general branches of human knowledge" and let's imagine D'Alembert asking someone $\mathcal{Q}(x)$: "draw me a branch of knowledge that deals with x", the same way we could have asked him: "give me a chapter of the Encyclopedia that deals with x".

If we have at our disposal a phylomemy of science $\phi \in \mathcal{R}$ over the temporal series of clustering $\mathcal{C}^*$ of roots $\mathcal{L}$, for all x in $\mathcal{L}$, we can draw inspiration from information retrieval to shape an answer to $\mathcal{Q}(x)$ and give d'Alembert a branch of science that will describe some contexts of x and their evolution.

Let's consider a representation $\varphi \in \mathcal{V}$ of ϕ made of a set of branches $\varphi = \bigcup_k B_k$ and let $\mathcal{C}_x$ be the set of all fields that containing x in $\mathcal{C} = \bigcup_{C^T \in \mathcal{C}^*} C^T$, $\mathcal{C}_x$ covering a set of periods $\mathcal{T}_x$; and $\mathcal{C}_{B_k}$ be the set of all the fields of the branch B_k. Then, an answer to D'Alembert's question could consist in pointing to a given phylomemetic branch B_k containing some elements of $\mathcal{C}_x$ and state: "here is the dynamical description of a domain of research concerned by x". Two metrics to assess the relevance of B_k to $\mathcal{Q}(x)$ are of interest at this stage:

- The *precision* $\xi_x^k = \frac{|\mathcal{C}_x \cap \mathcal{C}_{B_k|\mathcal{T}_x}|}{|\mathcal{C}_{B_k}|}$ of B_k against x. Within periods $\mathcal{T}_x$ where x is present in ϕ, it is the proportion of clusters of B_k that mention x (probability to observe x by choosing at random a cluster in B_k among the clusters from periods where x is present in ϕ),

- The *recall* $\rho_x^k = \frac{|C_x \cap C_{B_k}|}{|C_x|}$ of B_k against x. It is the proportion of clusters of C_x that are in the branch B_k (probability to be in B_k when choosing a cluster about x at random in ϕ).

An answer B_k to $Q(x)$ will be all the more precise regarding x that its precision ξ_x^k is high. But it will provide all the more information about the different historical contexts of x that its recall is high. Precision and recall are generally antagonistic and consequently, the questioner has to indicate the desired 'trade-off' in his/her question.

The quality of an answer $Q(x)$ can be assessed with respect to the desired level of trade-off $\lambda \in [0\ 1]$ between precision and recall with the following F-score function:

$$F_\lambda(x,k) = \frac{(1+f(\lambda)^2).(\xi_x^k.\rho_x^k)}{\rho_x^k + f(\lambda)^2.\xi_x^k}$$

where $f(\lambda) = tan(\frac{\pi.\lambda}{2})$. For $\lambda = 0$, only the precision counts, whereas for $\lambda = 1$[13], only the recall counts.

Several branches could mention x, which means that we also have to know which branch to propose first as an answer. For this reason, we introduce a generic choice function Ψ (a random variable to be determined later) that tells which branch among the branches containing x will be the first to be examined by D'Alembert.

The objective function for the evaluation of the relevance of ϕ in answering $Q(x)$ thus becomes:

$$F_\lambda^x(\varphi) = \Sigma_{B_k \in \varphi | B_k \cap C_x \neq \emptyset} \Psi_x(k).F_\lambda(x,k) \quad (1)$$

If we wanted to provide D'Alembert with independent answers for every x, we could search, for each question $Q(x)$, the phylomemetic network with the best F_λ^x score. However, this would prevent D'Alembert from having a global vision of science and of the articulation between its different branches: answers based on different queries will indeed not necessarily be comparable.

In order to provide a global representation φ of a domain of science delineated by $\mathcal{L}$, we thus have to assess the relevance of a particular phylomemetic network for its ability to answer any question $Q(x)$ D'Alembert might ask about elements of $\mathcal{L}$. Since some x may interest D'Alembert more than others, the optimal φ should take into account the interest profile of D'Alembert for elements of $\mathcal{L}$. We will call Ξ the choice function over $\mathcal{L}$ that determines the probability of D'Alembert asking for a particular x.

Given the choice functions Ξ and Ψ, the global F-score of a representation φ of a phylomemy can be defined as:

$$F_\lambda(\varphi) = \Sigma_{x \in \mathcal{L}} \Xi(x).F_\lambda^x(\varphi) \quad (2)$$

Note that Ξ is a property of the questioner whereas Ψ can be a property of either the respondent or the questioner. Ξ and Ψ are both random variables on which the meaning of a given phylomemy projection φ will depend. We will see in section 6 how these two functions can be determined empirically.

Branches with high recall for a given x will tend to be more complex to interpret because the contexts in which x are very varied and provide a huge amount of information, whereas branches with high precision for a given x will tend to be simpler because they target very homogeneous contexts.

Consequently, F_λ is a score of quality whose parameter λ can be related to a desired *level of observation*. For high λ values, the phylomemies with the highest F_λ score will be the ones that include large complex branches whereas for low λ values, phylomemies with small homogeneous branches will score the highest.

The objective function F_λ gives meaning to the problem of choosing a projector $\mathcal{R} \longmapsto \mathcal{V}$: given a level of observation of a phylomemy ϕ, what is the best projector to optimize the information conveyed by the corresponding phylomemetic network? The next section proposes an approach to solve this problem.

3.2.4 Adaptive inter-temporal matching and step phylomemetic networks

Let's consider a reconstruction operator ${}^4\Phi : \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(\mathcal{C}), w)^2$ that from a partial phylomemy reconstruction workflow (steps 1 to 3), reconstructs the inter-temporal matching links from a set of clusters. For any desired scale of observation λ of ϕ, we can evaluate any projection φ in $\mathcal{V}$ with $F_\lambda(\varphi)$.

As mentioned in 3.2.3, pioneer work on phylomemy reconstruction have so far only taken into account uniform projectors. Here, we propose to consider a new class of adaptive projectors that, as we will show, are naturally parameterized by the level of observation λ defined over $\mathcal{R}$, adapt to the internal dynamics of each branch of science and outperform uniform projectors.

[13] We consider for $F_\lambda(1)$ the limit value of F_λ when $f(\lambda) \to \infty$ which is ρ_x^k.

Definition: Let $\Pi : \mathcal{R} \longmapsto \mathcal{V}$ be a projector, $\phi \in \mathcal{R}$ a phylomemy and $\varphi = {}^4\Phi_\Pi(\phi) = \bigcup_k B_k$ a phylomemetic network. Let $\mathcal{C}_{B_k}$ be the set of clusters of C^* such as ${}^4\Phi_\Pi(\mathcal{C}_{B_k}) = B_k$. Π is a *uniform step projector* for ϕ if and only if:

$$\forall B_k,\ \exists \delta \in [0,1], \forall C \in \mathcal{C}_{B_k}, {}^4\Phi_\Pi(C) = {}^4\Phi_\delta(C)$$

Definition: φ is a *step phylomemetic network* if and only if there is a phylomemy ϕ and a *uniform step projector* $\Pi : \mathcal{R} \longmapsto \mathcal{V}$ such as $\varphi = {}^4\Phi_\Pi(\phi)$.

Step phylomemetic networks are projections of phylomemies of particular interest because the meaning of inter-temporal matching links within each of their branches is uniform, with a minimal strength of $\delta^\lambda_{B_k}$, a property that makes it easier to interpret their morphology. Moreover, the optimization of the objective function F_λ makes it possible to find the step phylomemetic network that fits best to the internal dynamics of each branch of science at a given level of observation. The family of step phylomemetic networks extends the family of phylomemetic networks obtained by mean of uniform projectors.

The *sea-level rise* algorithm described below generates, for a given phylomemy ϕ over a temporal series of clustering C^*, a family of step phylomemetic networks parameterized by the level of observation λ.

Let's start to notice that if $\lambda = 1$, then only the *recall* counts in F_λ, so that the larger the branches of φ, the better. Except in rare cases[14], this is achieved by setting $\delta = 0$ for inter-temporal matching such that in $\mathcal{V}$ the highest number of temporal links is retained. Thus, for $\lambda = 1$, the best projector is the uniform projector $\Pi = 0$ that results in the phylomemy projection $\varphi_0 \in \mathcal{V}$.

To estimate *locally*, for every cluster C^T and for any $\lambda \in [0\ 1]$, the most appropriate value of $\delta^\lambda_{C^T}$, we proceed by recurrence. We perform an adaptive "sea-level rise" with uniform projectors Π_δ within each subset of $\mathcal{C}^{B_l} \subset C^*$ defined by a phylomemetic branch B_k of φ_0. Beforehand, we will need the following definition :

Definition: splitting threshold of a phylomemetic branch in $\mathcal{R}$ (the first time B_k splits as a result of a rise of the inter-temporal matching threshold). Let $\varphi = {}^4\Phi_\Pi(\phi) = \bigcup_k B_k$ be a step phylomemetic network over the temporal series of clustering C^*. Let $\mathcal{C}^{B_k}$ and δ_{B_k} be such that ${}^4\Phi_{\delta_{B_k}}(\mathcal{C}^{B_k}) = B_k$. The splitting threshold $\beta(B_k) = \delta'_{B_k}$ of B_k is the smallest value $\delta'_{B_k} > \delta_{B_k}$ such as ${}^4\Phi_{\delta'_{B_k}}(\mathcal{C}^{B_k})$ has more than one connected components[15].

***Sea-level rise* algorithm over ϕ at level λ**

The *Sea-level rise* algorithm operates a recurrence on the first values where branches split after raising δ within each branch independently, coupled to a branching process (cf. Figure 4).

ALGORITHM

Initialization :

The starting point of the sea-level algorithm is the phylomemetic lattice $\varphi_0 =^4 \Phi_{\delta=0}(\phi)$ composed of branches $\{B^0_k\}_k$ with an internal inter-temporal threshold $\delta_{B^0_k} = 0$ for all branches. The quality of the initial phylomemetic network is $F_\lambda(\varphi_0)$ with the initial state described by $(\{(B^0_k, \delta_{B^0_k} = 0)\}_k, F_\lambda(\varphi_0))$. We will note $\overline{B}$ the fact that the branch B is *locked*.

Recurrence:

While there are unlocked branches:

1. Let the state be $(\{(B_k, \delta_{B_k})\}_k \cup \{(\overline{B}_l, \delta_{\overline{B}_l})\}_l, F_\lambda(\varphi))$ with $\varphi = \bigcup_k B_k \bigcup_l \overline{B}_l$. Choose the unlocked branch $B_u \in \{B_k\}_k$ that has the largest set of fields $\mathcal{C}^{B_u} \subset C^*$ and compute its splitting threshold $\delta'_{B_u} = \beta(B_u)$. If $\beta(B_u)$ does not exist, locks B_u as $\overline{B}_u$ and reiterate 1. from the state $(\{(B_k, \delta_{B_k})\}_{u \neq k} \cup \left((\overline{B}_u, \delta_{B_u}) \cup \{(\overline{B}_l, \delta_{\overline{B}_l})\}_l\right), F_\lambda(\varphi))$,

[14] In practice this is almost always true. We could build synthetic phylomemies in which not only recall would count for $\lambda = 1$, but these would not be very consistent with human activities.

[15] $\beta(B_k)$ might not exist if for example B_k is composed of a single cluster.

2. if $\beta(B_u) > 0$, with ${}^4\Phi_{\beta(B_u)}(C^{B_u}) = \bigcup_j B_u^j$, let $\varphi' = \bigcup_{u \neq k} B_k \bigcup_j B_u^j \bigcup_l \overline{B}_l$, then
   - If $F_\lambda(\varphi) \geq F_\lambda(\varphi')$ then *lock* the branch B_l that becomes and start again at 1) with the state $(\{(B_k, \delta_{B_k})\}_{u \neq k} \cup \left((\overline{B}_u, \delta_{B_u}) \cup \{(\overline{B}_l, \delta_{\overline{B}_l})\}_l\right), F_\lambda(\varphi))$,
   - If $F_\lambda(\varphi) < F_\lambda(\varphi')$ then start again at 1. from the state $\left(\{(B_k, \delta_{B_k})\}_{k \neq u} \cup \{(B_u^j, \delta'_{B_u})\}_j\right) \cup \{(\overline{B}_l, \delta_{\overline{B}_l})\}_l, F_\lambda(\varphi'))$ with $\varphi' = \bigcup_{u \neq k} B_k \bigcup_j B_u^j \bigcup_l \overline{B}_l$.

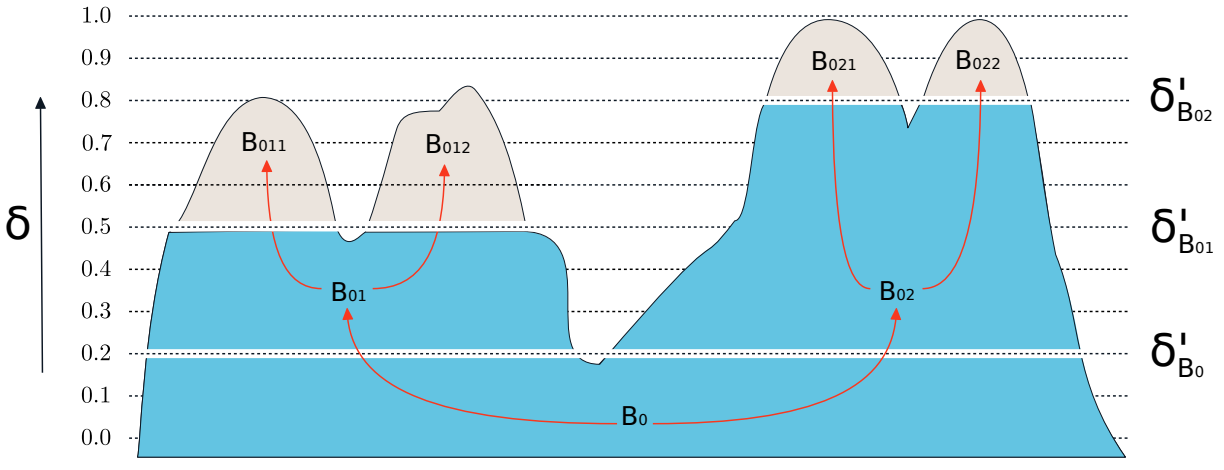

Figure 4: The elevation of similarity submerges the initial branch $\varphi_0 = B_0$ that first splits into two branches $B_{01} \cup B_{02}$ at δ'_{B_0} ; and then each of them splits at different thresholds δ'_{B_01} and δ'_{B_02} to create the final branches of $\varphi = B_{011} \cup B_{012} \cup B_{021} \cup B_{022}$.

The sea-level rise algorithm either locks a branch or splits a branch at each step. Since the maximum number of branches is bounded by the initial number of clusters in C^*, the number of steps of the algorithm is bounded by $card(C^*)$.

By design, the phylomemetic networks generated by this algorithm are *step phylomemetic networks*.

It is worth noticing that for a given branch $B_k \in \varphi_{\delta_0}$, raising δ in ${}^4\Phi_{\Pi_\delta}(C_{B_k})$ has for effect to disqualifies some of the inter-temporal matching links in the branch B_k of φ_{δ_0} and causes inter-temporal matching to be reprocessed. Most of the time, this leads to the pruning of links when no other combination of parents or children satisfy the new threshold, but sometimes this process can lead to the discovery of more distant links that pass the threshold and thus transform simple parentage into multiple parentage - or the reverse with longer temporal range.

Consequently, as it can be seen in Figure 3, even if they share the same set of clusters C, two observations of ϕ at different levels φ_λ and $\varphi_{\lambda'}$ ($\lambda \neq \lambda'$) will not necessarily share the same set of inter-temporal links between these clusters and it may not be possible to transform one into the other simply by pruning the links. Two phylomemetic lattices generated at two different level of observation λ and λ' might therefore convey very different information on the temporal structure of ϕ.

3.2.5 Phylomemetic networks and scale of observation

Branches of phylomemetic networks φ can have complex structures that can be rendered at different scales in $\mathcal{V}$ thanks to a synchronic merging of their clusters. This is particularly useful for visualization and numerous methods exist to represent this kind of multi-scale structure. A rather straightforward method in our case would be to merge all clusters that are similar enough within the same branch and the same period of time. Other methods like synchronic hierarchical clustering of clusters of branches can also be used.

But the natural clustering method in our case would be to extend the sea-level algorithm to $\mathcal{V}$ and to highlight the internal structures of phylomemetic branches. Raising the threshold δ within a given branch without recomputing the parents and children of clusters leads to the progressive pruning of links, to the disconnection of the temporal graph

that forms this branch and to the formation of several connected components until all the clusters are isolated. These connected components can then be used as a basis for a hierarchical synchronic clustering.

Definition: scales of observation in $\mathcal{V}$ of a phylomemetic branch.
Let B be a phylomemetic branch from of a step phylomemetic network with internal 'sea-level' δ_B (the weakest inter-temporal link has a strength of δ_B). Let $\theta :: \mathcal{V} \times [\delta_B\ 1] \longmapsto \mathcal{V} : (B, \delta) \rightarrow B'$ be the function that removes all the inter-temporal links of B that are strictly less than δ[16] . For $\delta > \delta_B$, $\theta(B, \delta) = \{B_l^\delta\}_{1 \leq l \leq s^\delta}$ is composed of s^δ connected components that form sub-branches of B. Moreover, s^δ is an increasing step function of δ with discontinuities at the values $S(B) = \{\delta^i\}_{1 \leq i \leq s_B}$ such as $s^{\delta^1} = 1$ is the number of sub-branches of $B = \theta(B, \delta_B)$ and $s^{\delta^{s_B}} \leq card(C_B)$ is the number of sub-branches of $\theta(B, 1)$. Finally, for $j > i$, the sub-branches $\{B_h^{\delta^j}\}_{1 \leq h \leq s^{\delta^j}}$ of $\theta(B, \delta^j)$ are by construction nested inside the sub-branches $\{B_l^{\delta^i}\}_{1 \leq l \leq s^{\delta^i}}$ of $\theta(B, \delta^i)$.

By construction, $\forall 1 \leq i \leq s_B, \forall C \in \mathcal{C}^T \cap C_B, \exists! l \in \{1..s^{\delta^i}\}\ such\ that\ C \in B_l^{\delta^i}$. Therefore, $\theta(B, \delta^i) \cap \mathcal{C}^T$ defines a non overlapping clustering over the fields of B at period T. Since $\{\theta(B, \delta^i)\}_{1 \leq i \leq s_B}$ are nested sets, the family of clusterings $\{\theta(B, \delta^i) \cap C^T\}_{1 \leq i \leq s_B}$ defines an endogenous synchronic hierarchical clustering over $C^T \cap B$ indexed endogenously by the scales of description $\{1, .., s_B\}$.

For a *level λ of observation* $\varphi_\lambda = \{B_k^\lambda\}_k \in \mathcal{V}$ of a phylomemy ϕ, for each phylomemetic branch B_k^λ, we can consequently define its endogenous *scales* of description $\{1, .., s_{B_k^\lambda}\}$ through the endogenous synchronic clustering of fields in B_k^λ and a choice for the merging procedure of the associated inter-temporal matching links $\{C^T \overset{\omega}{\longleftrightarrow} C'^T | (C^T, C'^T) \in \mathcal{C}^2\}$.

The merge of inter-temporal matching links can be formally described by the choice of an operator Ω (*e.g.* the weight of a set of merged links is the min/avg/max of their weights):

$$\mathcal{H}_k^\lambda : \{1..s_{B_k^\lambda}\} \times \mathcal{V} \longmapsto \mathcal{V} :$$

$$\begin{cases} (s, C^T \in B_k^\lambda) \longmapsto H_\lambda(C^T) \\ for, T \prec\prec T' \\ \{C^T \overset{w}{\longleftrightarrow} C'^{T'} \in \mathcal{C}^2 \times [0\ 1]\} \longmapsto \{H_\lambda(C^T) \overset{\Omega(B_k^\lambda, w)}{\longleftrightarrow} H_\lambda(C'^{T'}) \in H_\lambda(\mathcal{C})^2 \times [0\ 1]\} \end{cases} \tag{3}$$

An example of a branch described at several scales is given by Figure 5 where Ω is the function that keeps only the strongest links during a merge. The advantage of this definition of *scales* for phylomemetic branches is that it endogenously adapts to the internal complexity of each branch but nevertheless makes it possible to define a scale of description for a full phylomemetic network: at scale 1, all branches have at most one cluster at each period corresponding to the synchronic merge of all the clusters of the branch at that period. For scale $s > 1$, branches start branching according to their internal structure but remain in constant number.

3.2.6 Computing the ancestors beyond the time horizon

The inter-temporal matching procedure described in 3.2.4 may introduce artificial splits of phylomemetic branches for at least two reasons:

- **Horizon of observation.** Generally, the oldest period T_0 of a phylomemy is not the beginning of the story but the horizon of our capacity to observe textual production due to incomplete database. For example, it is well known that as of today, most abstracts of scientific papers are missing before 1990. This puts a limitation to the identification of phylomemetic branches by the sole use of ${}^4\Phi$ since two branches might well have a common ancestor before T_0 but appear unrelated in the phylomemy just because this ancestor is not observed. The ancestor is beyond our horizon of observation.

[16] In this operation, contrary to what is made in 3.2.4, inter-temporal matching is not reprocessed.

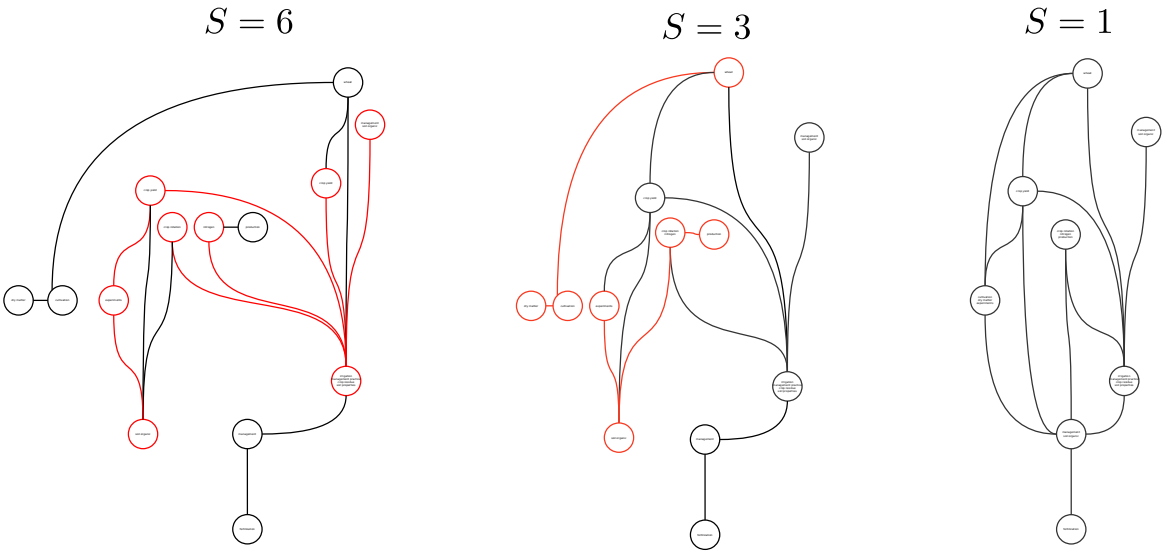

Figure 5: Endogenous scales of a branch. The branch *wheat management* of the glyphosate phylomemy of Figure 2 has six scales of description. Here we display its internal structure at scales 1, 3 and 6. Groups and links highlighted in red are affected by changes of scale: they have either appeared, merged or divided.

- **Fast emergence.** When a domain emerges quickly, it might be explored at the same time in several distinct directions that are so internally consistent that the closest parents are always in the same sub-branch from the very beginning. In that case, inter-temporal matching alone cannot detect that they are part of the same research question, regardless of the level of observation.

In both cases, we would like phylomemy reconstruction to be able to identify at some level of observation that two branches from the same phylomemetic network and sharing some common terms from their very beginning are related.

Ghost ancestors. To mitigate this limitation, we have added an additional step in the reconstruction workflow that takes place in $\mathcal{V}$ and consists in searching for common 'ghost' ancestors to emerging fields that have no parents (see Figure 6). Given a phylomemetic network $\varphi_\lambda = \{B_k^\lambda\}_k \in \mathcal{V}$ of a phylomemy ϕ, if a field $C^T \in C^T$ is the oldest field of a branch B^k with internal 'sea-level' δ_{B_k}, the similarities with the other fields of C^T are processed. A ghost ancestor between C^T and C'^T is then created at the period immediately preceding T every time $\Delta(C^T, C'^T) \geq \delta_{B_k}$. This ghost ancestor is linked to both C^T and C'^T with weight $\Delta(C^T, C'^T)$. This operation can cause branches to merge into $\mathcal{V}$ through ghost ancestors. We note ${}^4\overline{\Phi} : \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(\mathcal{C}), w)^2$ the reconstruction operator that includes the processing of the ghost ancestors. For the sake of clarity, when there is no ambiguity, we will continue to call *branch* the set of phylomemetic branches linked to the same ghost ancestor. The quality F_λ will be computed on ${}^4\overline{\Phi}_\Pi(\mathcal{C})$.

4 Results

We have implemented the inter-temporal reconstruction workflow ${}^4\overline{\Phi}$ as a module of the free software *Gargantext* [17] [28] that already implements ${}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$ (see the details of the implementation in SI Appendix D).

We have evaluated the quality of the full workflow $\phi = {}^4\overline{\Phi} \circ {}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$ from the point of view of F_λ compared to results obtained by [17], as well as its capacity to provide an accurate description of the evolution of scientific domains through qualitative evaluation. The perspectives that this new methodology offers for the interaction with large set of documents through visualization are further detailed in [45].

Thereafter, we will now consider for Ξ and Ψ the choices functions that seem the most consensual to us without any prior knowledge:

[17] Gargantext is a text-mining software under GNU aGPL Licence written in haskell and purescript. See http://gargantext.org.

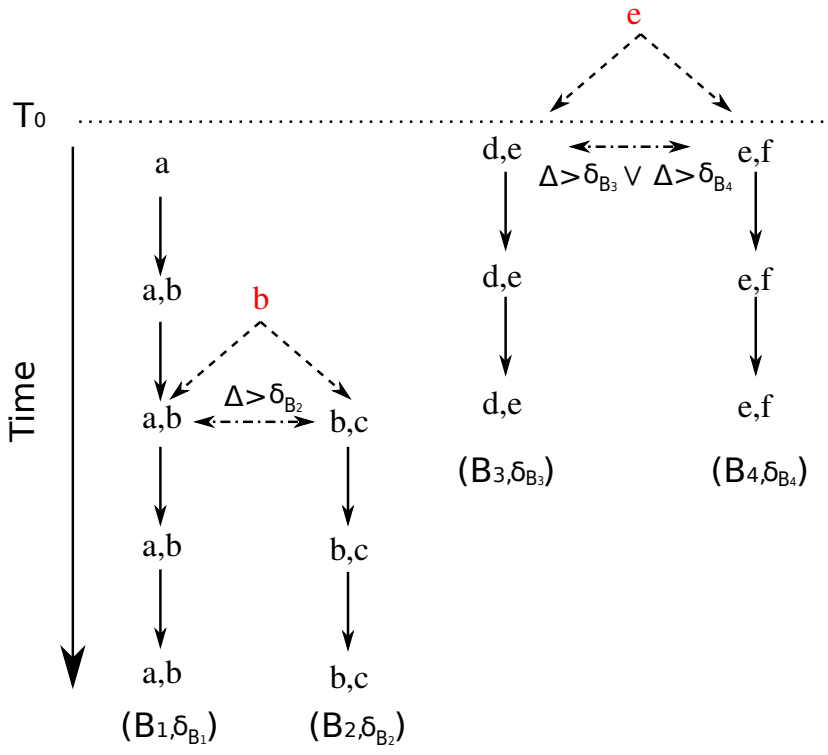

Figure 6: Reconstruction of ghost ancestors. When a field $C^T \in B_k$ has no parent, the similarities with the other fields of C^T are processed and a ghost ancestor is created every time this similarity is higher than δ_{B_k}. This operation can cause branches to merge.

- $\Xi(x, \_)$ is a random variable which chooses terms in $\mathcal{L}$ with a uniform probability,
- $\Psi(x, \_)$ is a random variable which chooses a branch $B_k \in \mathcal{B}_x$ with a probability proportional to its number of fields.

We will see, in section 6, how Ξ and Ψ can be empirically determined from specific uses and research questions.

Quality function. Given the choices of Ξ and Ψ, the objective function $F_\lambda(\varphi)$ on $\varphi = \{B_k\}_k$ can be written with Equation 4:

$$F_\lambda(\varphi) = \Sigma_{x \in \mathcal{L}} \frac{1}{|\mathcal{L}|} . \Sigma_{B_k \in \mathcal{B}_k^x} \frac{|B_k|}{\Sigma_{B_j \in \mathcal{B}_k^x} |B_j|} . F_\lambda(x, k) \tag{4}$$

where: $\mathcal{B}_k^x = \{B_k | B_k \cap C_x \neq \emptyset\}$

4.1 Comparison with previous work

We have compared our method to previous work on several case studies. We will now highlight the main parameters of the workflow $\phi = {}^4\overline{\Phi} \circ {}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$ that are required to interpret the phylomemetic networks of these case studies.

4.1.1 Workflow settings

Corpora and key-phrases extraction ($^1\phi$). We have chosen several distinct corpora to test the robustness of our results and to illustrate the wide range of applications of our method:

- *Domain specific academic literature.*

  - **Glyphostate literature** (see Appendix C for a qualitative analysis). A corpus of 16,8k documents retrieved in the Web of Science (WoS) and PubMed on September 2020 with the query *glyphosate OR roundup OR "round-up"* and merged into a single corpus with duplicates removed. The analysis of this specialized corpus highlights the diversity of current research on glyphosate and the environmental and public health issues induced by its uses. We have named this corpus $\mathcal{D}_{glyphosate}$.
  - **Quantum computing literature**. A corpus of 29k documents retrieved from the WoS with the query *("quantum computer" OR "quantum computing" OR "quantum processing" OR "quantum algorithm" OR "quantum communication")* on July 4th 2019. The analysis of this specialized corpus highlights the birth of a new scientific domain that emerged during the 1990s. We have named this corpus $\mathcal{D}_{quantum}$.

- *Interdisciplinary academic literature.* We have chosen a corpus of 6000 top-maincited papers extracted from the WoS on October 2020 and for which at least one of the authors is affiliated to the National Center for Scientific Research (CNRS). CNRS being an interdisciplinary research organism, this corpus is by construction highly interdisciplinary and contains a limited number of documents in each discipline, yet, the phylomemy reconstruction gives it possible to have an overview of the main research streams and highlights the way they combine (see [45] for a detailed description). We have named this corpus $\mathcal{D}_{CNRS}$.

- *Generic time-stamped corpora.* To illustrate our method's ability to reconstruct knowledge dynamics form any kind of time-stamped corpus as well as to process very short documents, we have applied our method to the descriptions of 6k arms of clinical trials related to the Covid-19 and published between 2020-02 and 2021-01. These descriptions have been dated by week of deposit in the WHO database and $\mathcal{L}$ as been defined as the list of all the treatments considered in these descriptions. The phylomemy reconstruction highlights the different research paths and discoveries made around the Codiv-19 outbreak and could as such become a useful tool for worldwide coordination between researchers (cf. SI C.3). We have named this corpus $\mathcal{D}_{CT}$.

For each of these corpora, we have extracted the core vocabulary using Gargantext[18]. This core vocabulary has been further processed to compute the set of *roots* for each case study and their temporal co-occurrences. The corpora and the lists of terms necessary for the reproduction of this study are downloadable from archive [60].

Similarity measures ($^2\Phi$). In these case studies, we have adopted the *confidence* similarity measure[19] which belongs to the class of syntagmatic similarity measures (cf. SI A). This similarity measure has proven to be a good indicator of the hypernymy / hyponymy relations between terms [29]. It is also very easy to interpret for non-specialists. Phylomemies that use such measures are reflecting actual interactions between terms and clusters in the phylomemy are generally made up of the core vocabulary of some academic sub-community.

Community detection or clustering ($^3\Phi$). The strictest notion of cluster (or community) in a graph is the notion of clique, *i.e.* subsets whose elements are all linked to one another. Cliques are a fundamental and impassable level of granularity for clustering algorithms. Most clustering methods extend the notion of a cluster by relaxing the notion of a clique. Since phylomemy reconstruction generates an endogenous hierarchical clustering over a temporal series of clustering (cf. 3.2.5) we can afford to be as parsimonious as possible on the initial assumptions we make about clusters' properties and adopt cliques as the fundamental granularity of our reconstruction. But all cliques are not equally relevant for the mapping of research domains, and cliques of small size (*e.g.* 3 or 4) or with weak internal links in most cases provide but little information and introduce some noise in the reconstruction. Moreover, keeping these weak cliques increases the number of cliques exponentially and therefore generally renders the algorithms intractable. We consequently filter the less relevant cliques according to the parameters presented in Table 3 and Table 4. In what follows, we will use both the *maximal cliques* and the *frequent item sets* (Fis) methods (see definitions and details in SI D.3).

[18] Gargantext allows an interactive selection of terms assisted by machine learning

[19] The *confidence* between two terms i and j is the max of the estimation of the two probabilities of having one term given the presence of the other in the same contextual unit.

Table 3
Parameters for filtering cliques in each case study. Cliques of small size or with weak internal links are filtered out. The *coverage* column indicates the proportion of roots from $\mathcal{L}$ that are in at least one clique in the corresponding filtered clique set $\mathcal{C}$.

corpus	minimal clique size	links threshold	coverage
$\mathcal{D}_{glyphosate}$	7	0.001	84%
$\mathcal{D}_{quantum}$	4	0.001	80%
$\mathcal{D}_{CNRS}$	5	0.001	90%
$\mathcal{D}_{CT}$	1	0	100 %

Table 4
Parameters for filtering Fis in each case study. Fis of small size or of small support are filtered out. The *coverage* column indicates the proportion of roots from $\mathcal{L}$ that are in at least one Fis in the corresponding filtered $\mathcal{C}$.

corpus	minimal Fis size	minimal Fis support	coverage
$\mathcal{D}_{glyphosate}$	7	5	11%
$\mathcal{D}_{quantum}$	4	3	52%
$\mathcal{D}_{CNRS}$	5	2	49%
$\mathcal{D}_{CT}$	1	1	100 %

Inter-temporal matching ($^4\Phi$). The Jaccard similarity measure is commonly used as a means to compare sets of elements of the same nature, making this measure a good candidate for inter-temporal matching. However, elements of clusters in phylomemies (*roots*) are not necessarily comparable in their usages (*e.g.* hyperonymes vs. hyponymes) ; which is why we have introduced a variant of the Jaccard measure to weight the contributions of *roots* according to their characteristics and to take into account this specificity of language. Starting with a simple adaptation, we've weighted the contributions of terms according to their frequency in the corpora through a sensibility parameter $\sigma \in [0\ 1]$. The chosen function Δ^σ is a standard Jaccard similarity for $\sigma = 0$, puts more weight on rare specific terms for $\sigma > 0$ and more weight on frequent generic terms for $\sigma < 0$:

$$\begin{cases} \Delta :: [-1\ 1] \times \mathcal{P}(\mathcal{L}) \times \mathcal{P}(\mathcal{L}) \longmapsto [0\ 1] \\ \Delta^\sigma(C, C') = 0 \ if \ C \cap C' = \varnothing \\ \Delta^\sigma(C, C') = 1 \ if \ C = C' \\ \Delta^\sigma(C, C') = \dfrac{|C \cap C'|}{|C \cup C'|} \ if \ \sigma = 0; \\ \Delta^\sigma(C, C') = \dfrac{\Sigma_{\{p_i | i \in \cap(C,C')} \frac{1}{log(g(\sigma)+p_i)}}{\Sigma_{\{p_i | i \in \cup(C,C')} \frac{1}{log(g(\sigma)+p_i)}} \ if \ \sigma > 0; \\ \Delta^\sigma(C, C') = \dfrac{\Sigma_{\{p_i | i \in \cap(C,C')} log(g(\sigma) + p_i)}{\Sigma_{\{p_i | i \in \cup(C,C')} log(g(\sigma) + p_i)} \ if \ \sigma < 0; \\ ; \end{cases} \tag{5}$$

where $g(\sigma) = \frac{1}{tan(\frac{2}{pi} * \sigma)}$

Phylomemies reconstructed for values of σ close to 1 will therefore tend to highlight the evolution of specific sub-domains in which clusters sharing hyponyms are likely to be related, while phylomemies reconstructed for values of σ close to -1 will tend to highlight the evolution of very generic domains where clusters sharing hyperonyms are likely to be related. The quality of these reconstructions can then be compared using F_λ.

4.1.2 Quantitative evaluation

We will now compare the quality of the phylomemies obtained with the sea-level rise algorithm and the best quality that can be achieved with a uniform projector taken in $\{\pi_\delta | \delta \in \{0, 0.1, ..., 0.9, 1\}\}$. We will use FIS to build series of temporal clustering $\mathcal{C}^*$ based on parameters taken from Table 4.

Since by definition, uniform projectors are step phylomemetic projectors, the best step phylomemetic network is necessarily of higher quality than the best uniform phylomemetic network. However, finding the step phylomemetic network that optimizes F_λ is done in a rugged quality landscape and is therefore a difficult task that deserves a paper in itself. The algorithm proposed in this paper can undoubtedly be improved but, as we will show, it is a good starting point.

As can be seen on Figure 7, for the various case studies and for most levels of observation λ, the sea level algorithm outperforms or is at least as good as the original method. Red dots on these figures highlight the values of λ for which we can improve our implementation of the sea-level rise algorithm and soften the influence of the various cliques' filters[20]. Moreover, SI D.6 demonstrates that for an alternative objective function, the sea-level rise algorithm outperforms uniform step projectors all the time.

These results proves that the sea-level rise algorithm succeeds in adapting locally to the internal dynamics of the branches and in producing better precision and recall couples. Step phylomemetic networks obtained by this new algorithm should therefore be preferred over networks obtained by uniform projectors.

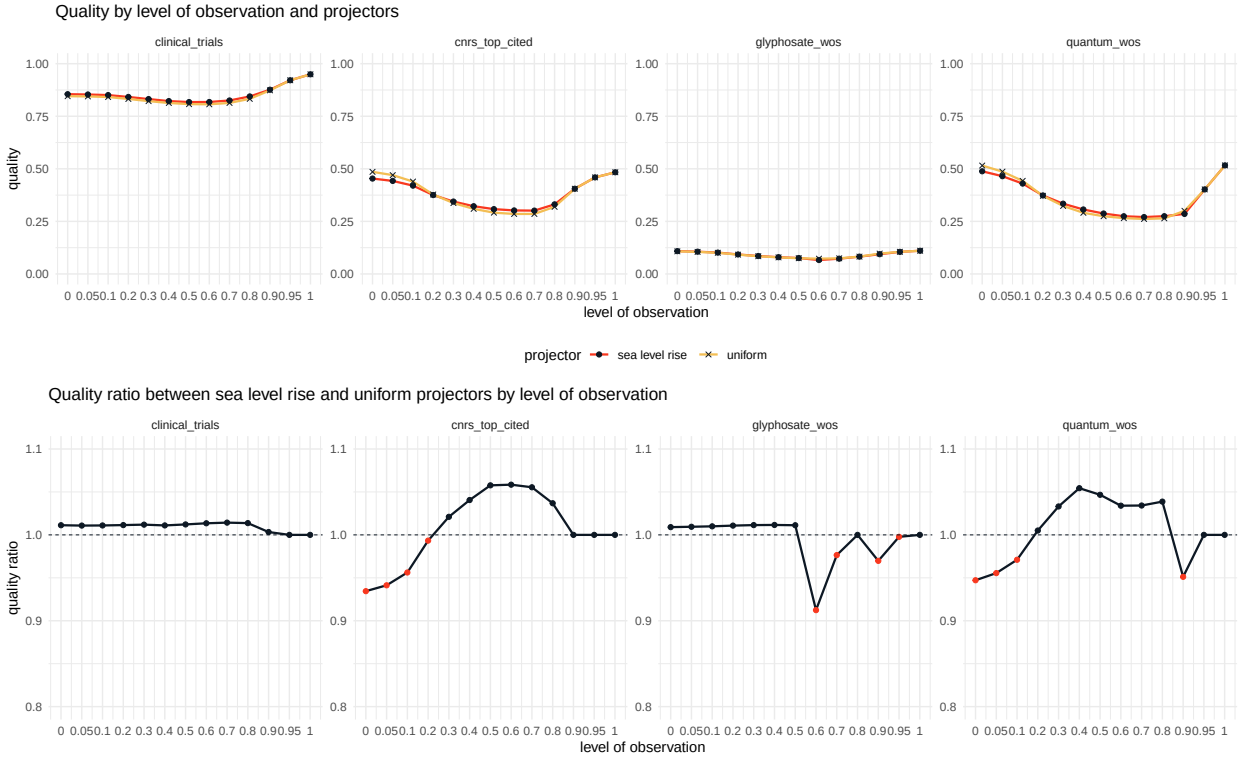

Figure 7: Comparison between the sea-level rise and uniform projectors for four distinct corpora.

4.2 External and qualitative validation

In order to assess the ability of phylomemies to fit and extend scholars expertise about knowledge dynamics of their fields, we compared in details the results of its application to two case studies: *glyphosate research* and research on *science mapping and visualization*.

4.2.1 40 years of glyphosate research

The field of glyphosate-related research (delimited by the corpus $\mathcal{D}_{glyphosate}$) is particularly interesting to illustrate the relevance of our method. Glyphosate has been introduced as a pesticide in the mid seventies and by the 1990s had become the most marketed herbicide active ingredient, a position it has more or less held ever since [61]. Literature

[20] By definition, we can always find a suitable step phylomemetic network able to include any uniform step projector (see 3.2.4). Red dots are therefore evidences of unsatisfactory optimizations in the way we locally increment and elect δ (see SI D.5)

on glyphosate is quite recent, most of it being digitized, and the knowledge it contains is of great importance from a health and economic point of view.

The specificity of glyphosate-related literature also lies in the fact that there is yet no consensual synthesis of the knowledge it contains. Glyphosate is a controversial herbicide about which literature reviews and historical analyses are regularly published, some emphasizing the advantages of glyphosate (e.g. [31]) or the absence of associated risks ([26, 73]), others reviewing the risks for health and environment [61, 46, 56] and issues related to the the emergence of herbicide-resistant weeds [48].

In such scientific context, there is a high risk of *selection bias* when selecting publications for literature reviews and monographs, *i.e.* to 'cherry-pick' the publications that would confirm certain theoretical claims over others. In such situation, as highlighted by [18], phylomemy reconstruction could be useful to give an overall picture of the field, as objective as possible inasmuch as it would include as many publications on the topic as possible and process them equally, solely on the basis of a definition of what constitutes a valid publication. Such phylomemy could highlight, before any further considerations of what is important and what is not, the main issues addressed by the scientific community and their global trends.

In order to identify the main issues of research on glyphosate as objectively as possible, we have made a synthesis of key historical reviews written by glyphosate proponents and skeptics [31, 2, 61, 35, 46, 56].

To summarize, there has been an evolution in the questions addressed by glyphosate literature, from research on its effects on plants, its benefits for agriculture and its usages (1990-2005, a period called the 'golden age' of glyphosate) to the problem of herbicide-resistant weeds management (2005 -), the assessment of environment and health issues for animals (2004 -) and the question of its threats to humans health (2014 -). These different research questions emerge clearly from synthesizing the structure of the various literature reviews under analysis (cf. SI section C.1).

The analysis of the phylomemy of glyphosate research at different scales of observation (cf. Figure 2 and SI. figures 17 to 19) makes it possible to successfully identify these different research questions, the details of their ramifications and their development. Full details of this analysis are provided in SI. C.

4.3 Dynamical state-of-the-art of literature related to science and knowledge mapping

The phylomemy of the knowledge dynamics corpus $\mathcal{D}_{maps}$ analyzed in the in-depth state-of-the art of this paper (cf. SI B) at level of observation 0.3 is presented on Figure 8.

We can observe that this phylomemy correctly describes our state-of-the-art in its temporal dimension:

- The pioneer field of *citation analysis* was predominant during the 1970's (branches no.1) before passing the baton to what will become the core of *bibliometry* and *scientometry* in the early 1990's (branches no.3).
- In parallel, *co-word* and *co-occurrence* analysis [63, 11] emerged in the mid 1980's (branch 2) and enjoyed a revival of interest in the middle of the 2000's (branch 4) as a result of the ICT revolution. Our paper belongs to this more recent branch.
- In the mid 2000's, the field of *information retrieval* developed topic modeling methods (branch 6) that were subsequently applied to digital libraries and text-classification (branch 7) as well as social media analysis (branch 8).
- At the same time, the long established field of *concept mapping* found concrete applications in the domains of *education* and *learning process* (branches 5).

5 Discussion

5.1 Limits and continuous improvement

Phenomenological reconstruction ($\mathcal{O} \longmapsto \mathcal{R} \longmapsto \mathcal{V}$) can lead to a misunderstanding or a biased representation of an object $O \in \mathcal{O}$ for several reasons. First, some important observables for the understanding of O could have been neglected or inadequately measured in the process $\mathcal{O} \longmapsto \mathcal{R}$. Regarding the reconstruction of knowledge dynamics, this bias can be expected to diminish over time as text-mining techniques improve and as an increasing proportion of knowledge production contexts produces ever more structured and accessible digitized traces.

Second, since by definition, dimension reduction reduces the number of variables under consideration, some important information for the understanding of R could be lost in $\mathcal{R} \longmapsto \mathcal{V}$ (typically, two elements that are distant or unrelated in $\mathcal{R}$ could appear arbitrarily close after being projected in $\mathcal{V}$, see [23] for a good example).

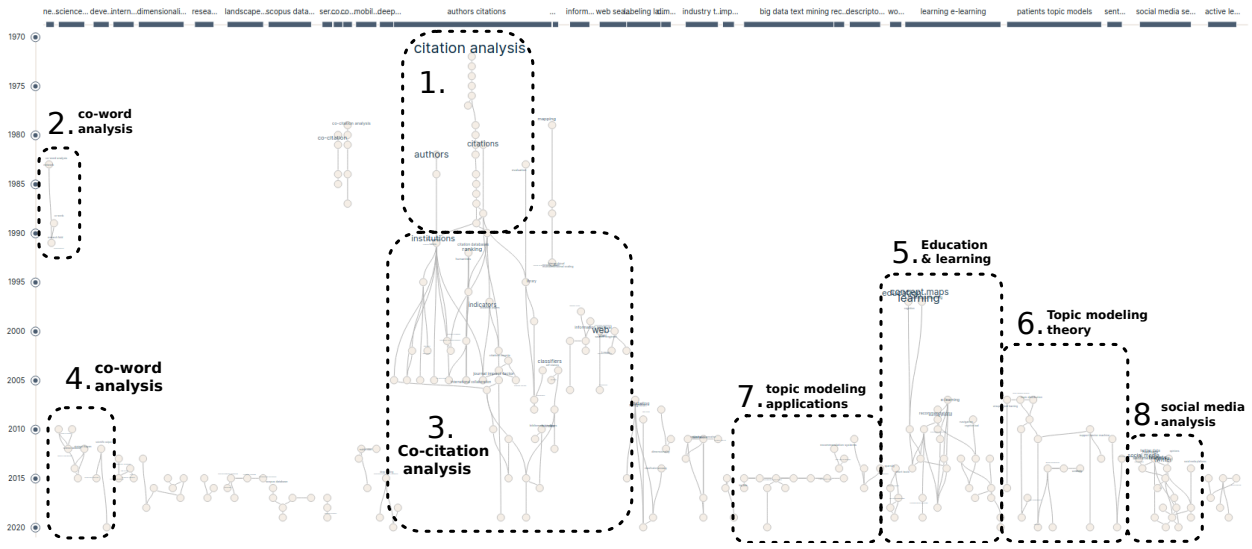

Figure 8: Phylomemetic network of the literature related to science and knowledge mapping ($\mathcal{D}_{maps}$) at level 0.3 with branches smaller than 3 filtered out. Dashed boxes have been manually annotated. An interactive version of this phylomemetic network available at `http://maps.gargantext.org/phylo/knowledge_visualization/memiescape` and can be downloaded from the archive [59].

These potential limitations are important to keep in mind. In the same way as different 2D projections of a world map provide complementary information about Earth's geography (some projections conserve angles, other areas, etc.), different methods for phenomenological reconstruction are undoubtedly needed to fully grasp a body of knowledge. This point was already highlighted by D'Alembert [32]:

> "Knowledge is impossible to draw as a whole in a truthful manner, but only through the choice of a point of view that is both arbitrary and inevitable [...] One can create as many different systems of human knowledge as there are world maps having different projections, and each one of these systems might even have some particular advantage possessed by none of the others."

We have adopted in this paper what we think is the most generic approach to knowledge dynamics: the mining of unstructured textual data. When other meta-data become available, this approach will definitely benefit from being complemented or hybridized with other approaches such as citation analysis and co-author analysis.

The objective function described by Equation 2 can also be applied to assess alternative phylomemy reconstruction workflows, a process that will lead to a continuous improvement of the phenomenological reconstructions and can also lead to local adaptations to the different contexts of knowledge production.

For example, we have compared in preliminary explorations the role of the inter-temporal matching function Δ via a sensibility analysis on the parameter of our measure Δ^{σ} (cf. Fig.9). The variation of σ from -1 to 1 reinforces the weight of generic to specific kinship connections and thus influences the topology of the resulting branches[21]. We also have explored the role of cliques detection and filtering algorithm with the comparison between maximal cliques and frequent item sets algorithms[22] (by using the parameters of Table 4 and Table 3). The results of Fig. 10 encourage us to review in depth what the suitable conditions for using either algorithms at best would be, according to the nature of corpus or the targeted level of observation.

Beyond their assessment with regard to the optimization of F_{λ} from a quantitative perspective, *frequent item sets* and *max cliques* can also be compared from a qualitative perspective. *Frequent item sets* method relies on a notion of *support* (*i.e.* the number of documents represented by a given cluster) and makes it possible to follow through time the presence of papers mentioning a particular combination of terms. On the other hand, the *max cliques* technique will

[21] In Fig.9, with $\mathcal{D}_{CNRS}$, the use of $\sigma = 1$ increases the number of branches by 20 compared to $\sigma = -1$ and thus improves the quality of the phylomemy for low value of λ, i.e, when high precision is required.

[22] The maximal clique algorithm [4] finds all the largest cliques in a graph ; the frequent item set algorithm find all sets C such as all elements of C appear together in at least one document.

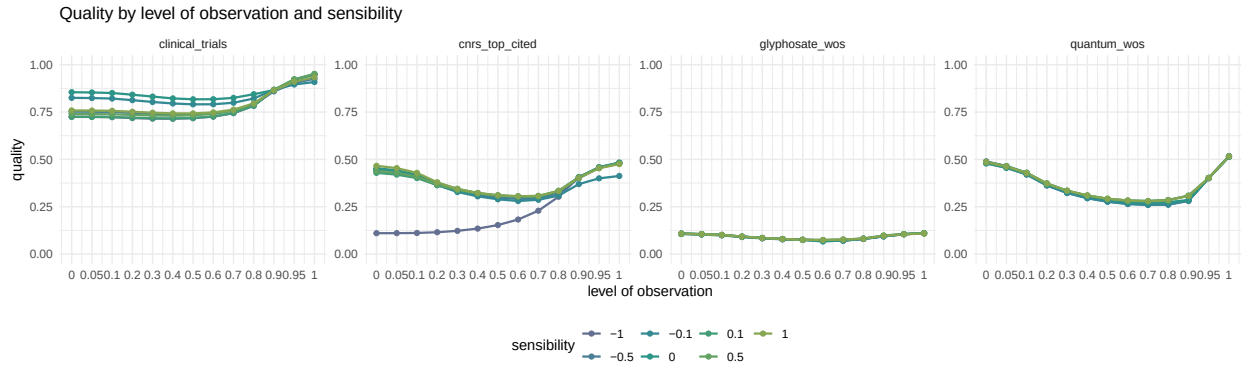

Figure 9: Quality comparison by a variation of sensibility for four distinct corpora.

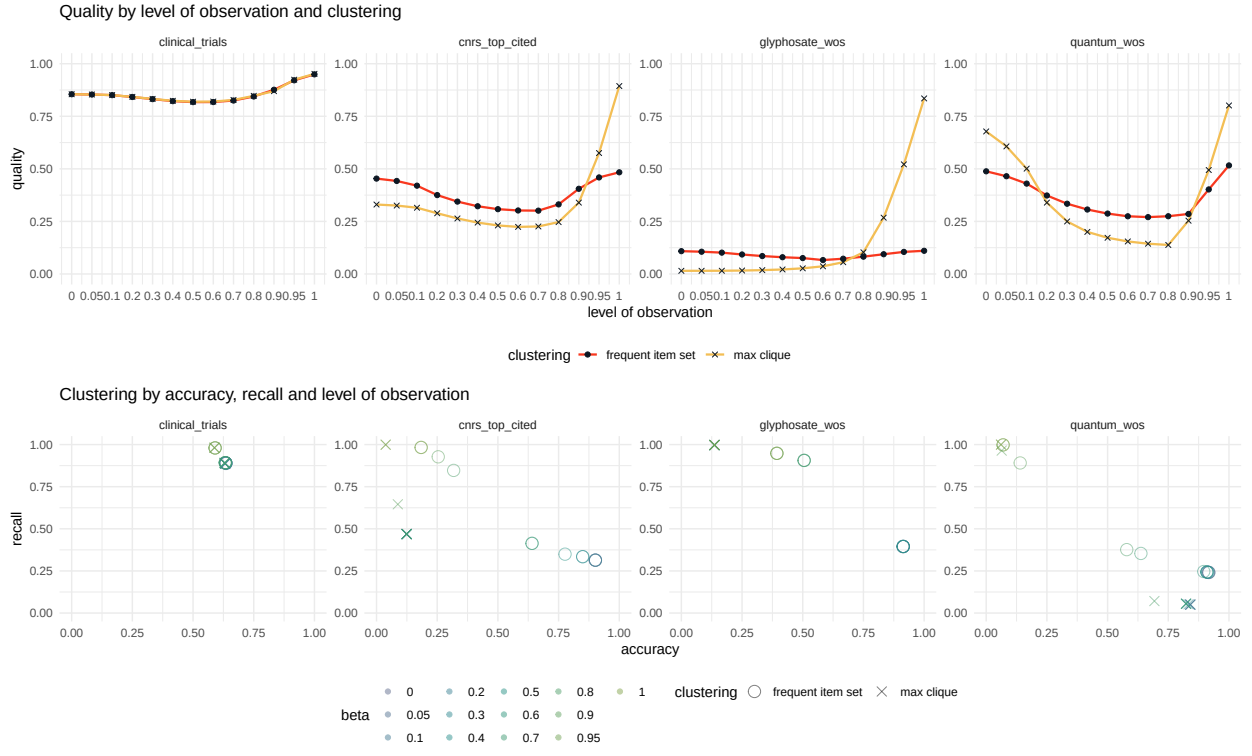

Figure 10: Comparison of performance between the frequent item set and the max clique techniques for four distinct corpora.

allow to detect cliques of terms that are not explicitly associated together in a single article, thus making it possible to detect emerging domains before they are consolidated through publications. Consequently, each approach produces different phylomemetic shapes that might answer different research questions.

Exploring different reconstruction workflows is consequently not only about improving performance. It also involves comparing the respective contributions of each option to the characterization, at different level and scales of observation, of the case studied. Going back to D'Alembert's analogy mentioned in 1, the phylomemy reconstruction can be seen as an exploration tool or as a telescope where each operator of the whole workflow is a slot designed to embed specific lenses; and choosing a suitable lens remains the prerogative of the analyst.

5.2 Embodied cognition and users' preferences

Although it is common practice to take users' preferences into account in information retrieval tasks (cf. [30, 54] for examples), it is much less common in the literature on science and knowledge mapping (left part of SI Fig 12). One reason could be that this field of science aims to provide as objective a view as possible of a whole scientific landscape, with the hope that we can capture some basic truth about what that landscape is. We must not forget, moreover, that part of this literature has its origins in scientometrics, one of the goals of which was to assess scientific production. This evaluation should therefore not depend on the evaluator.

In this paper, we propose a third way between the temptation to reach an absolute ground truth and the *ad-hoc* adaptation to a particular user's preferences.

The present operation of reconstruction acknowledges that knowledge dynamics dwell in a very high-dimensional space $\mathcal{R}$ whose which elements require to be projected in a lower-dimensional space $\mathcal{V}$ in order to be grasped by the human mind. It also takes place at the level of a collective representation of a body of knowledge: once the perimeter of the representation has been defined via a corpus and a vocabulary ($\mathcal{D}$ and $\mathcal{L}$), the aim is to find a representation that can be common to any question formulated on the basis of this vocabulary by a community of users. These constraints lead us to make the distinction between two classes of adaptation to the user's preferences[23]:

- The choices of a level and a scale allows users to agree on the intrinsic and extrinsic complexity of the representation they want to see,
- The choice of functions Ψ and Ξ determines the aspects of the knowledge domain on which the collective interest is focused and on which reconstruction should be the most accurate.

Consequently, while the level and scale should be considered as tunable parameters allowing users to interactively explore an object, Ψ and Ξ should be viewed as parameters to be learned by the system in a semi-supervised way, in order to maximize its relevance as a coordination tool for a collective. Ψ models the order in which users inspect the branches of knowledge when several of them are relevant to answer a given question $x \in \mathcal{L}$. Ξ models the frequency at which each question x is asked by a community of users. These two functions have a significant impact on the reconstruction of phylomemetic networks, as documented in SI D.6, where an alternative formulation for Ξ has been tested.

If we imagine that phylomemy reconstruction is used by a community of users to either retrieve documents or collectively assess the shapes and properties of a research landscape, then Ψ and Ξ can be estimated and revised according to the collective behavior of users. In this perspective, we cannot consider the representations provided by the systems as external to users anymore. We rather have to see these representation as co-constructed through the users' interactions with a digital environment.

In this perspective, phenomenological reconstruction, as formalized in this paper by the chain $\mathcal{O} \longmapsto \mathcal{R} \longmapsto \mathcal{V}$, is envisioned as an interface with the digital world. It is a fundamental step in the elaboration of meaning. But it is itself influenced by the meaning we give to things, encapsulated in the functions Ξ and Ψ.

We consequently have a circular dependency between the preferences of the members of a collective that uses a device and the parameters of this same device. This circular dependency can be related to Francisco Varela's conception of cognition [65], thought as the result of the sensory-motor interactions of a living being with his environment. In this sense, our approach is in line with his epistemology of embodied cognition [66]. Collective cognition emerges from morphogenetic and path-dependent processes in our interactions with our environment. By simplifying reality by the transformations $\mathcal{O} \longmapsto \mathcal{R} \longmapsto \mathcal{V}$, it allows us to grasp the complexity of structures of $\mathcal{O}$ despite our limited cognitive capacities. Meaning emerges in those circular interactions, there is no meaning or "ground truth" outside these processes.

Ultimately, we think that knowledge maps - and in particular those that take time into account -, when implemented in certain interactive and adaptive digital devices, can embody a form of collective representation of the collective knowledge produced by a community (e.g. the academic world). These representations can play, at the collective level, a role analogous to individual mental representations in the theories of embodied action, and can become true tools for the coordination of representations and actions among individuals.

[23] In principle, our entire methodology could also be applied to a single user asking for the best answer to a particular query independently of potential other queries, but this is not where its originality lies.

6 Conclusions and perspectives

In this paper, we have set a general framework for a phenomenological reconstruction of science and knowledge dynamics from large digitized data sets.

We then have extended previous works, and in particular phylomemy reconstruction introduced by [16] in several ways:

- we have formalized the notion of *level* and *scale* of knowledge dynamics as complex systems,
- we have proposed a new class of meaning for the reconstruction of knowledge dynamics formalized by a new objective function parameterized by the level of observation,
- we have properly formalized the concept of phylomemy as distinct from the concept of phylomemetic networks,
- we have proposed a new reconstruction algorithm for phylomemetic networks reconstruction that outperforms previous ones thanks to a new objective function,
- we have shown in case studies that this approach produces representations of knowledge dynamics close to the ones that can be obtained by synthesizing the points of view of experts on a given domain (glyphosate literature and knowledge and science mapping literature),
- we have demonstrated with cases studies that this approach can be applied to any kind of unstructured corpora, even on relatively small data sets or short texts,
- we have proposed a new temporal clustering on dynamical graphs that is naturally part of the process of multi-level and multi-scale reconstruction of phylomemies,
- we have integrated user preferences into our framework by providing an interaction model and contextualizing the different elements of our reconstruction workflow in the theoretical framework of the embodied cognition,
- By applying our method to the state-of-the-art of this paper, we have illustrated how it could be systematically applied to generate such extended historical analysis which might be helpful to give more context to the scope and contributions of scientific papers.

The method proposed in this paper is moreover fully generic and can be applied to any kind of unstructured corpora containing some form of collective knowledge, from social media posts to patents, as it has been demonstrated for example on the WHO COVID-19 clinical trials database (cf. SI. C.3).

The diversity of our case studies demonstrates that the proposed methodology makes it possible to address with the same methodology a wide variety of textual contents, from big data to small data, and from short texts to long texts, where other approaches are more focused on specific types of data. This paves the way for cross-comparisons between different knowledge production arenas via a unified methodology (*e.g.* academic literature, patents, news, (micro-)blogs, etc.).

The clear distinction between the notion of levels and scales in phenomenological reconstruction of knowledge dynamics might be generalized to other phenomenological reconstruction approaches. A dedicated interactive visualization allowing the user to navigate between scales and interact with knowledge dynamics at a given level of observation has been developed and is presented in details in [45].

Again, we are not claiming that the proposed approach should be used on an exclusive basis. The present reconstruction would definitely gain to be hybridized with other phenomenological reconstruction methods to offer different points of view to the analyst.

So, to answer the Little Prince, phylomemy reconstruction is a good start to produce an outline of science for him to color as he wishes with complementary approaches.

Data availability and replication

All data and code used for this paper have been published in open access for full reproducibility :

- Data and supplementary material are available on the archive [60],

- A dedicated software for interactive visualization is available on the archive [59] and presented in [45].
- The haskell code developped for generating the phylomemetic networks has been integrated to the Gargantext software and is available at https://gitlab.iscpif.fr/gargantext/haskell-gargantext in the branch dev-phylo (forthcoming merge with the master branch).

Acknowledgments

This research was supported by the Complex Systems Institute of Paris Île-de-France (https://iscpif.fr), the *EPIQUE* project (ANR-16-CE38-0002-01), the ANR FORCCAST project and the EU FuturICT 2.0 project. We warmly thank Bruno Gaume for his fruitful comments on our work.

References

[1] Abell, S.K., Lederman, N.G. (Eds.), 2007. Handbook of research on science education. Lawrence Erlbaum Associates, Mahwah, N.J.

[2] Benbrook, C.M., 2016. Trends in glyphosate herbicide use in the United States and globally. Environmental Sciences Europe 28, 3. URL: https://doi.org/10.1186/s12302-016-0070-0, doi:10.1186/s12302-016-0070-0.

[3] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022.

[4] Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M., 1999. The maximum clique problem, in: Handbook of combinatorial optimization. Springer, pp. 1–74.

[5] Bourgine, P., Brodu, N., Deffuant, G., Kapoula, Z., Müller, J.P., Peyreiras, N., 2009. Formal epistemology, experimentation, machine learning, in: HAL Archives Ouvertes. Chavalarias et al.. https://hal.archives-ouvertes.fr/hal-00392486, pp. 10–14.

[6] Boyack, K.W., Klavans, R., 2010. 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 61, 2389–2404. URL: http://doi.wiley.com/10.1002/asi.21419, doi:10.1002/asi.21419.

[7] Braam, R.R., Moed, H.F., van Raan, A.F.J., 1991. Mapping of science by combined co-citation and word analysis. I. Structural aspects. Journal of the American Society for Information Science 42, 233–251. doi:10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I. 00352 bibtex: braamMapping1991.

[8] Börner, K., 2010. Atlas of science: Visualizing what we know. The MIT Press.

[9] Börner, K., 2015. Atlas of knowledge: anyone can map. The MIT Press, Cambridge, Massachusetts. Bibtex: bornerAtlas2015.

[10] Börner, K., Chen, C.M., Boyack, K.W., 2003. Visualizing knowledge domains. Annual Review of Information Science and Technology 37, 179–255. doi:10.1002/aris.1440370106. 00868 WOS:000179918000006 bibtex: bornerVisualizing2003.

[11] Callon, M., Courtial, J., Turner, W., Bauin, S., 1983. From Translations to Problematic Networks - an Introduction to Co-Word Analysis. Social Science Information Sur Les Sciences Sociales 22, 191–235. doi:10.1177/053901883022002003. 00548 WOS:A1983QW30800003 bibtex: callonTranslations1983.

[12] Callon, M., Rip, A., Law, J., 1986. Mapping the dynamics of science and technology: Sociology of science in the real world. Springer.

[13] Cambe, J., Grauwin, S., Flandrin, P., Jensen, P., 2020. Exploring and comparing temporal clustering methods. arXiv:2012.01287 [physics] URL: http://arxiv.org/abs/2012.01287. arXiv: 2012.01287.

[14] Chavalarias, D., 2020. From inert matter to the global society - Life as multi-level networks of processes. Philosophical Transactions of the Royal Society B: Biological Sciences URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2018.0348, doi:http://dx.doi.org/10.1098/rstb.2019.0329.

[15] Chavalarias, D., Cointet, J.P., 2008. Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. Scientometrics 75, 37–50. URL: http://link.springer.com/10.1007/s11192-007-1825-6, doi:10.1007/s11192-007-1825-6. 00030 bibtex: chavalariasBottomup2008.

[16] Chavalarias, D., Cointet, J.P., 2013a. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. PloS one 8, e54847. URL: http://dx.plos.org/10.1371/journal.pone.0054847. 00000 bibtex: chavalariasPhylomemetic2013.

[17] Chavalarias, D., Cointet, J.P., 2013b. Science Phylomemy, in: Places and Spaces, Places and Spaces. URL: http://scimaps.org/maps/map/science_phylomemy_159/, doi:http://scimaps.org/maps/map/science_phylomemy_159/. http://scimaps.org/maps/map/science_phylomemy_159/ bibtex: chavalariasScience2013.

[18] Chavalarias, D., Huneman, P., Racovski, T., 2020. Contribution of phylomemies to the understanding of the dynamics of science, in: The Dynamics of Science: Computational Frontiers in History and Philosophy of Science. Pittsburgh University Press, Pittsburgh.

[19] Chavalarias, D., Jean-Philippe, C., Duong, T.K., Cornilleau, L., Villard, L., Roth, C., Savy, T., 2011. Streams of Media Issues - Monitoring World Food Security, in: UN GLobal Pulse Symposium, United Nations, New York. URL: https://www.unglobalpulse.org/project/monitoring-food-security-issues-through-news-media-2011/.

[20] Chen, B., Tsutsui, S., Ding, Y., Ma, F., 2017. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. Journal of Informetrics 11, 1175–1189. URL: http://www.sciencedirect.com/science/article/pii/S1751157717300536, doi:10.1016/j.joi.2017.10.003.

[21] Chen, C., 2017. Science Mapping: A Systematic Review of the Literature. Journal of Data and Information Science 2, 1–40. URL: https://content.sciendo.com/view/journals/jdis/2/2/article-p1.xml, doi:10.1515/jdis-2017-0006. publisher: Sciendo Section: Journal of Data and Information Science.

[22] Chen, C., Song, M., 2019. Visualizing a field of research: A methodology of systematic scientometric reviews. PloS one 14.

[23] Chuang, J., Ramage, D., Manning, C., Heer, J., 2012. Interpretation and trust: designing model-driven visualizations for text analysis, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, Austin, Texas, USA. pp. 443–452. URL: https://doi.org/10.1145/2207676.2207738, doi:10.1145/2207676.2207738.

[24] Claveau, F., Gingras, Y., 2016. Macrodynamics of economics: A bibliometric history. History of Political Economy 48, 551–592. Publisher: Duke University Press.

[25] Cointet, J.P., Chavalarias, D., 2008. Multi-level Science mapping with asymmetric co-occurrence analysis: Methodology and case study. Networks and Heterogeneous Media , 267–276Bibtex: chavalariasMultilevel2008.

[26] Crump, K., 2020. The Potential Effects of Recall Bias and Selection Bias on the Epidemiological Evidence for the Carcinogenicity of Glyphosate. Risk Analysis 40, 696–704. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13440, doi:https://doi.org/10.1111/risa.13440. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.13440.

[27] Cui, Weiwei and Liu, Shixia and Tan, Li and Shi, Conglei and Song, Yangqiu and Gao, Zekai and Qu, Huamin and Tong, Xin, 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text, in: IEEE Transactions on Visualization and Computer Graphics, IEEE Educational Activities Department. pp. 2412–2421. doi:10.1109/TVCG.2011.239.

[28] Delanoë, A., Chavalarias, D., 2020. Mining the digital society - Gargantext, a macroscope for collaborative analysis and exploration of textual corpora. Forthcoming .

[29] Dias, G., Mukelov, R., Cleuziou, G., 2008. Mapping General-Specific Noun Relationships to WordNet Hypernym/Hyponym Relations, in: Knowledge Engineering: Practice and Patterns. Springer, pp. 198–212. 00003 bibtex: dias2008mapping.

[30] Druck, G., Mann, G., McCallum, A., 2008. Learning from labeled features using generalized expectation criteria, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 595–602.

[31] Duke, S.O., 2018. The history and current status of glyphosate. Pest Management Science 74, 1027–1034. doi:10.1002/ps.4652.

[32] d'Alembert, J.l.R., 1751. Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers. volume Tome 1. 00110 = "dictionary" bibtex: dalembertEncyclopedie1751.

[33] Fortunato, S., 2010. Community detection in graphs. Physics Reports 486, 75–174. URL: http://www.sciencedirect.com/science/article/pii/S0370157309002841, doi:10.1016/j.physrep.2009.11.002.

[34] Garfield, E., 1972. Citation analysis as a tool in journal evaluation. Science 178, 471–479.

[35] Gillezeau, C., van Gerwen, M., Shaffer, R.M., Rana, I., Zhang, L., Sheppard, L., Taioli, E., 2019. The evidence of human exposure to glyphosate: a review. Environmental Health 18, 2. URL: https://doi.org/10.1186/s12940-018-0435-5, doi:10.1186/s12940-018-0435-5.

[36] Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M., 2009. Topic Evolution in a Stream of Documents, in: Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. Proceedings, pp. 859–870. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611972795.74, doi:10.1137/1.9781611972795.74.

[37] Hull, D.L., 1988. Science as a process: an evolutionary account of the social and conceptual development of science. Science and its conceptual foundations, University of Chicago Press, Chicago. Bibtex: hullScience1988.

[38] Jo, Y., Hopcroft, J.E., Lagoze, C., 2011. The web of topics: discovering the topology of topic evolution in a corpus, in: Proceedings of the 20th international conference on World wide web, pp. 257–266.

[39] Jähnichen, P., Wenzel, F., Kloft, M., Mandt, S., 2018. Scalable Generalized Dynamic Topic Models. arXiv:1803.07868 [cs, stat] URL: http://arxiv.org/abs/1803.07868. arXiv: 1803.07868.

[40] Kessler, M., 1963. Bibliographic Coupling Between Scientific Papers. American Documentation 14, 10–&. doi:10.1002/asi.5090140103. 01294 WOS:A19632554A00006 bibtex: kesslerBibliographic1963.

[41] Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46, 604–632. Publisher: ACM New York, NY, USA.

[42] Li, G., Ge, W., Zhang, J., Kwauk, M., 2005. Multi-Scale Compromise and Multi-Level Correlation in Complex Systems. Chemical Engineering Research and Design 83, 574–582. URL: http://www.sciencedirect.com/science/article/pii/S0263876205727350, doi:10.1205/cherd.05093. publisher: Elsevier.

[43] Liao, D.p., Qian, Y.t., 2019. Paper evolution graph: multi-view structural retrieval for academic literature. Frontiers of Information Technology & Electronic Engineering 20, 187–205. URL: https://doi.org/10.1631/FITEE.1700105, doi:10.1631/FITEE.1700105.

[44] Lin, C., He, Y., 2009. Joint sentiment/topic model for sentiment analysis, in: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375–384.

[45] Lobbé, Q., Delanoë, A., Chavalarias, D., forthcoming 2021. Exploring, browsing and interacting with the multi-scale structures of knowledge .

[46] Martinez, D.A., Loening, U.E., Graham, M.C., 2018. Impacts of glyphosate-based herbicides on disease resistance and health of crops: a review. Environmental Sciences Europe 30, 2. URL: https://doi.org/10.1186/s12302-018-0131-7, doi:10.1186/s12302-018-0131-7.

[47] Millar, J.R., Peterson, G.L., Mendenhall, M.J., 2009. Document clustering and visualization with latent dirichlet allocation and self-organizing maps, in: Twenty-Second International FLAIRS Conference.

[48] Nandula, V.K., Reddy, K.N., Duke, S.O., Poston, D.H., 2005. Glyphosate-Resistant Weeds: Current Status and Future Outlook. Outlooks on Pest Management 16, 183–187. URL: http://openurl.ingenta.com/content/xref?genre=article&issn=1743-1026&volume=16&issue=4&spage=183, doi:10.1564/16aug11.

[49] Palla, G., Barabási, A.L., Vicsek, T., 2007. Quantifying social group evolution. Nature 446, 664–667. URL: http://www.nature.com.gate3.inist.fr/nature/journal/v446/n7136/full/nature05670.html, doi:10.1038/nature05670. 01073.

[50] Palmucci, A., Liao, H., Napoletano, A., Zaccaria, A., 2019. Where is your field going? A Machine Learning approach to study the relative motion of the domains of Physics. arXiv:1911.02890 [physics] URL: http://arxiv.org/abs/1911.02890. arXiv: 1911.02890.

[51] Rossetti, G., Cazabet, R., 2018. Community Discovery in Dynamic Networks: a Survey. ACM Computing Surveys 51, 1–37. URL:

http://arxiv.org/abs/1707.03186, doi:10.1145/3172867. arXiv: 1707.03186.

[52] Roth, C., Cointet, J.P., 2010. Social and semantic coevolution in knowledge networks. Social Networks 32, 16–29. URL: http://www.sciencedirect.com/science/article/pii/S0378873309000215, doi:10.1016/j.socnet.2009.04.005. bibtex: rothSocial2010.

[53] Rule, A., Cointet, J.P., Bearman, P.S., 2015. Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. Proceedings of the National Academy of Sciences , 201512221URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1512221112, doi:10.1073/pnas.1512221112. 00000.

[54] Shahaf, D., Guestrin, C., Horvitz, E., 2012. Trains of thought: Generating information maps, in: Proceedings of the 21st international conference on World Wide Web, pp. 899–908.

[55] Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., Leskovec, J., 2013. Information cartography: creating zoomable, large-scale maps of information, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 1097–1105. Bibtex: shahaf2013information.

[56] Singh, S., Kumar, V., Datta, S., Wani, A.B., Dhanjal, D.S., Romero, R., Singh, J., 2020. Glyphosate uptake, translocation, resistance emergence in crops, analytical monitoring, toxicity and degradation: a review. Environmental Chemistry Letters 18, 663–702. URL: https://doi.org/10.1007/s10311-020-00969-z, doi:10.1007/s10311-020-00969-z.

[57] 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, 265–269.

[58] Small, H., 1997. Update on science mapping: Creating large document spaces. Scientometrics 38, 275–293.

[59] (submitted), A., 2021a. Replication data : Exploring, browsing and interacting with multi-scale structures of knowledge. URL: https://doi.org/10.7910/DVN/WLI9B5, doi:10.7910/DVN/WLI9B5.

[60] (submitted), A., 2021b. Replication Data for: Draw me Science - multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. URL: https://doi.org/10.7910/DVN/SBH3EI, doi:10.7910/DVN/SBH3EI.

[61] Székács, A., Darvas, B., 2012. Forty Years with Glyphosate, in: Hasaneen, M.N.A.E.G. (Ed.), Herbicides-properties, synthesis and control of weeds. intechopen ed., pp. p. 247–284.

[62] Tacchella, A., Napoletano, A., Pietronero, L., 2020. The Language of Innovation. PLOS ONE 15, e0230107. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230107, doi:10.1371/journal.pone.0230107. publisher: Public Library of Science.

[63] Terzopoulos, D., 1985. Co-occurrence analysis of speech waveforms. IEEE transactions on acoustics, speech, and signal processing 33, 5–30.

[64] Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., Jain, A., 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95. URL: http://www.nature.com/articles/s41586-019-1335-8, doi:10.1038/s41586-019-1335-8.

[65] Varela, F.J., 1979. Principles of biological autonomy. The North Holland series in general systems research ; 2, North Holland, New York.

[66] Varela, F.J., Thompson, E., Rosch, E., 2000. The embodied mind: cognitive science and human experience. 8. print ed., MIT Press, Cambridge, Mass. OCLC: 249281910.

[67] Wang, C., Blei, D., Heckerman, D., 2015. Continuous Time Dynamic Topic Models. arXiv:1206.3298 [cs, stat] URL: http://arxiv.org/abs/1206.3298. arXiv: 1206.3298.

[68] Wang, C., Blei, D.M., 2011. Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, ACM Press, San Diego, California, USA. p. 448. URL: http://dl.acm.org/citation.cfm?doid=2020408.2020480, doi:10.1145/2020408.2020480.

[69] Wang, X., Cheng, Q., Lu, W., 2014. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. Scientometrics 101, 1253–1271. URL: https://link.springer.com/article/10.1007/s11192-014-1347-y, doi:10.1007/s11192-014-1347-y.

[70] Weeds, J., Weir, D., 2005. Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. Computational Linguistics 31, 439–475. URL: https://doi.org/10.1162/089120105775299122, doi:10.1162/089120105775299122. publisher: MIT Press.

[71] Wei, X., Croft, W.B., 2006. 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.

[72] White, H.D., McCain, K.W., 1998. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of the American Society for Information Science 49, 327–355. URL: http://doi.wiley.com/10.1002/%28SICI%291097-4571%2819980401%2949%3A4%3C327%3A%3AAID-ASI4%3E3.0.CO%3B2-W, doi:10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-W.

[73] Williams, G.M., Aardema, M., Acquavella, J., Berry, S.C., Brusick, D., Burns, M.M., Camargo, J.L.V.d., Garabrant, D., Greim, H.A., Kier, L.D., Kirkland, D.J., Marsh, G., Solomon, K.R., Sorahan, T., Roberts, A., Weed, D.L., 2016. A review of the carcinogenic potential of glyphosate by four independent expert panels and comparison to the IARC assessment. Critical Reviews in Toxicology 46, 3–20. URL: https://doi.org/10.1080/10408444.2016.1214677, doi:10.1080/10408444.2016.1214677. publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10408444.2016.1214677.

[74] Yang, Y., Yao, Q., Qu, H., 2017. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. Visual Informatics 1, 40–47.

[75] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., Stanley, H.E., 2017. The science of science: From the perspective of complex systems. Physics Reports 714-715, 1–73. URL: http://www.sciencedirect.com/science/article/pii/S0370157317303289, doi:10.1016/j.physrep.2017.10.001.

SUPPORTING INFORMATION FOR "DRAW ME SCIENCE - MULTI-LEVEL AND MULTI-SCALE RECONSTRUCTION OF KNOWLEDGE DYNAMICS WITH PHYLOMEMIES"

Authors

A Phenomenological and theoretical reconstruction

When facing complex systems, phenomenological reconstruction is an essential step in the modeling process (cf. Figure 11). Phenomenological reconstruction provides both inspiration for new theoretical reconstructions and tests for existing competing theoretical reconstructions. Theoretical reconstruction provides some indication of the objects to be examined in phenomenological reconstruction and predicts some patterns that might be found in the latter.

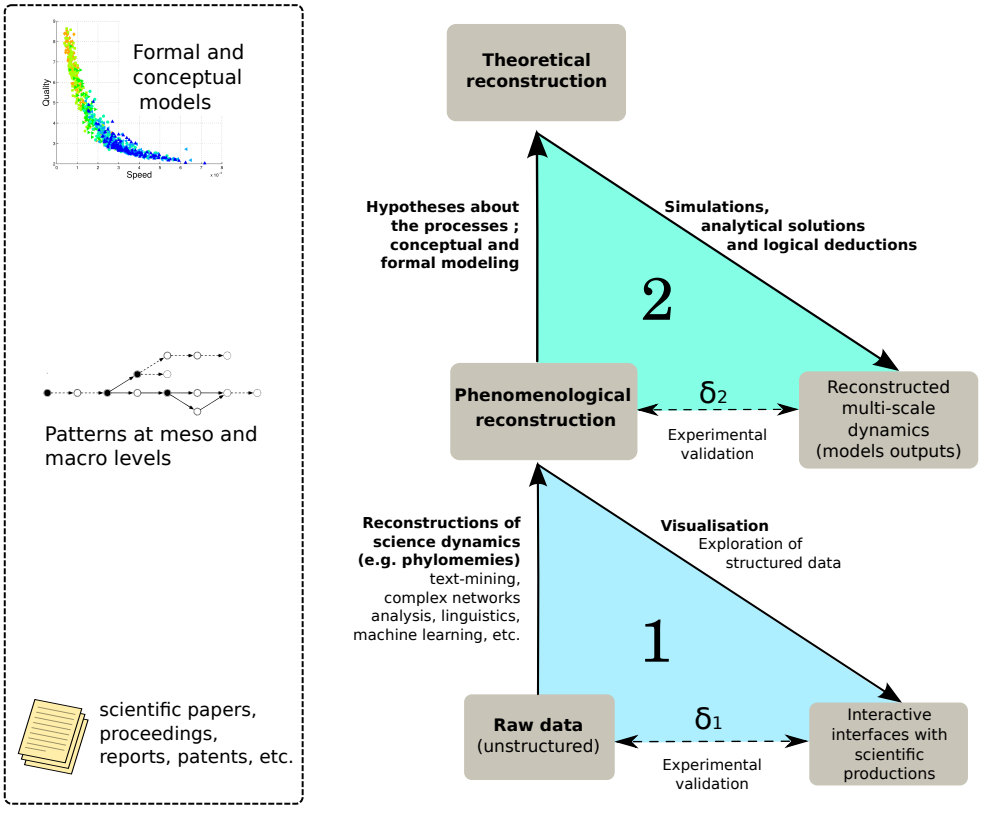

Figure 11: Articulation between phenomenological reconstruction and theoretical reconstruction in the case of the study of science dynamics. Source:

B In-depth state of the art

B.1 Research areas related to science and knowledge mapping

In order to identify the main areas concerned by the *science map* and *knowledge map* research areas we have performed a co-word analysis on a corpora $\mathcal{D}_{maps}$ of 14,374 documents retrieved on 2020-04-27 in the Web of Science with the query :

"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 "mapping research" OR "visualization of knowledge" 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")).

This analysis has been made with the text-mining software Gargantext (http://gargantext.org, free software). Its main outputs are the first order map based on syntagmatic relations between terms (cf. Fig. 13), that highlight domains produced by cohesive communities with homogeneous vocabulary ; and a second order map based on paradigmatic relations between terms (cf. Fig. 12), that highlights domains that focus on similar scientific objects. Interactive versions of these maps are available on http://maps.gargantext.org/maps/sciencemaps and a stand alone version can be downloaded from the digital archive .

Fig. 12 reveals the three major scientific orientations that structure this literature:

- The *bibliometrics* approach (on the left) aims at the analysis of large corpora of publications in order to qualify the underlying socio-semantic structures,
- The *information retrieval and documents classification* approach (on the right) aims at finding optimal methods to interact with large corpora of publications,
- The *education science and cognition* approach (at the top) focuses on methods that enhance the capacities of learners.

These three main approaches to automatic corpus analysis are, for some of them, divided into subfields of research with different histories, often carried out by distinct communities :

1. *Bibliometrics* has two major components, clearly distinct on fig. 13:
   - **Citation and co-citation analysis** The field of *scientometrics* emerged in the 1970s with the aim of measuring and analyzing scholarly literature. Primarily focused on *citation analysis* and the assessment of scientific output, it quickly diversified with the analysis of large scientific landscapes through methods such as *co-citation* and *bibliographic coupling* . Later on, following the creation of the *Web*, these methods have been generalized to bibliometrics with the emergence of *hyperlinked data* . Methods to describe the *conceptual structures* of science such as *research fronts*, *hot topics* and *trends*, etc. , sometimes studied together as socio-semantic networks , have come at the forefront of this research domain. Over the last decade, a growing number of contributions have proposed temporal reconstructions of the citation landscape .
   - **Co-word analysis** is a bottom-up approach first developed by sociologists in the 1980s to reconstruct the dynamics of *research themes* out of words *co-occurrence*, has developed to a generic approach to map knowledge dynamics in unstructured corpora . Citation and co-word analysis share common objectives, namely to understand the structures of science, which is reflected by their proximity on the map of Figure 12 where they form a whole that covers what is commonly called *science of science studies*. Their parallel development quickly paved the way to hybrid research between co-word and citation analysis . Nowadays, *science maps* and knowledge maps are interdisciplinary objects of research that result from both quali-quantitative and socio-technical processes . The growth of scientific databases has finally stimulated the visualization of wide *citation landscapes* or complex atlas of sciences .
2. **Latent semantic analysis and topic modeling** became a very active area of research in the early 2000s at the instigation of a community of statisticians who proposed the *latent dirichlet allocation* method as a way to

characterize *collections of documents*. Despite being focused on *document classification* , *recommendation* or *sentiment analysis* , parts of their most recent works have started to investigate the field of mapping scientific issues and their dynamics. The idea of working on latent semantic spaces has thereafter been taken up by the machine learning community with approaches such as words embedding . These approaches belong to the same class of language processing (dealing with the paradigmatic axis of language) and appear in the same cluster on the map of SI Figure 13.

3. **Concept and semantic maps.** In the 1990s, both mapping techniques gained entry to the field of *science of education* as a way to carry out *knowledge integration* . Influenced by psychology and cognition concerns, a *concept map* can be defined as a graphical representation designed to highlight the relationships between ideas or *key concepts* . It thus aims to clarify a given topic as well as its underlying *cognitive structure* by means of *ontologies*, *mind map*, *mental models*, etc. Unlike co-words approaches, *concept maps* were initially supposed to translate elements of knowledge built out of the top-down choices of *learners* and *teachers*. 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* .

Other research domains are not focused on developing knowledge representation methodologies but borrow existing techniques from these four sub-fields to study their own objects of research. Among these peripheral domains, the fields of *knowledge management* , *business intelligence* and *patent analysis* stand out and are clearly visible in a fifth cluster on SI Figure 13.

Overall, the main research communities interested in the question of knowledge dynamics are, on the one hand, bibliometrics, with the distinct but sometimes hybridized approaches of co-citation analysis and co-term analysis, and, on the other hand, the field of information retrieval. Finally, some recent research in the field of machine learning and neural networks has also addressed the issue of document classification and knowledge representation.

B.2 Criteria for comparing methodologies

In order to highlight the advantages and disadvantages of the various methods found in literature for mapping knowledge or science evolution, we have identified some of their main characteristics (summarized in Table 1 and Table 2) :

- **Focus.** Part of the scientific literature and knowledge mapping focuses on the mapping of corpora covering one of several domain of science, which is necessary if one wants, as d'Alembert put it, "to distinguish the general branches of human knowledge, the points that separate or unite them"; another part, stemming from the domain of information retrieval, focuses rather on the representation of knowledge covered by documents associated with a specific query. These differences in initial objectives generally lead to very different reconstruction strategies. We will call the first category 'corpora based' and the second category 'query based'.

- **Units of analysis.** A first class of methods starts with the clustering of document to produce a documents space and then extract semantic information from the analysis of the content or meta-data of these documents ; a second class of methods directly work on semantic space reconstruction by either working on key-words or classifications (*e.g.* PACS or JEL codes) present meta-data or extracting terms directly from the content of documents (title, abstract or document's body) by way of text-mining. Methods of the first type generally belong to the field of information retrieval and document classification ; methods of the second type generally belong to the field of bibliometrics and science of science.

- **Identification of fields of knowledge.** Various methods for the identification of fields of knowledge, commonly described as structured or unstructured set of keywords, have been proposed. Some of them pose some constraints on the type of temporal reconstruction that can be performed:

  - *Co-occurence clusters (COC).* Fields of knowledge are defined as clusters in a graph of items (keywords, terms, codes, authors, etc.) where the links between items represent their similarity or distance according to a specific statistics on their co-occurrences.
  - *Topic modeled as words distribution (TM).* Fields of knowledge are defined as a probability distribution on terms appearance in a set of documents. This definition of topics is adopted in LDA and the general branch of topic modeling,

  - *Metagraph factorization (MGF).* Fields of knowledge are identified as clusters in multipartite graphs where the dimensions of the graph include content information, authors, citation data, etc. This approach consequently needs well structured documents to be applied,
  - *Clusters in citation networks (CCN).* Fields of knowledge are identified as clusters in a graph of documents where links are based on citation data (direct citation, co-citation, bibliographic coupling, etc.). These clusters are thereafter enriched with some labels extracted from the content of their documents. This approach consequently needs well structured documents to be applied.
  - *Embedding and clusterisation (WE).* Since pioneer work of , neural network have been increasingly used to embed items occurring in documents in high dimensional metric spaces. When this embedding is made time-dependent, these metric space can be reduced to 2D or 3D spaces where terms trajectories can analyzed or further clusterisation can take place that produces hierarchical dynamical maps of science .

- **Advanced Text-mining ?** (y/n): Some papers include advanced text-mining in their analysis (multi-gram extraction, synonyms detection, n-gram grouping, etc.), other papers simply consider every single monogram as a token to analyze knowledge dynamics. Obviously, the first type of approaches leads to much more meaningful maps.

- **Work on unstructured text**(y/n): Some methods require the presence of specific metadata in order to be applied: citation data, keyword data, etc. Others require only the presence of a string of characters in each document.

- **Work on short texts**(y/n): For some methods, the relevance of their production deteriorates sharply when the number of terms per document falls below a certain threshold. This is particularly the case for LDA-type methods, which use statistical characterization of documents based on word distribution. If the number of words in the document is not high enough, this statistical characterization loses precision and the quality of the result is diminished.

- **Work on 'small' data.**(y/n): Quantitative analysis of digitized corpora is based on the extraction of statistical regularities in the data and therefore requires a sufficient volume of data. However, the volume of data required to obtain a good representation varies considerably from one method to another. While co-word analysis can provide acceptable representation on corpora of a few thousand documents, methods based on word embedding and machine learning will require much more ($> 100{,}000$) to produce acceptable results.

- **Unconstrained number of topics**(y/n): Some methods, and in particular the initial formulation of LDA, require the number of topics to be an a priori knowledge for the reconstruction. This is an important limitation for these methods, as this parameter has a strong impact on the final results and their interpretation. Setting a fixed number of topics without any further information is necessarily arbitrary and there is no reason why each period of time should have the same number of scientific fields, regardless of the resolution of the observation.

- **Evolving topics** (y/n): Some temporal analysis methods require not only that a fixed number of topics per period be pre-defined, but also that the description of these topics be the same for any period. This is obviously an important limitation since the evolution of the subjects is one of the main pieces of information that temporal phenomenological reconstruction could provide.

- **Re-emerging topics** (y/n): An important aspect of temporal phenomenological reconstruction is to detect re-emerging topics, that might have disappeared for a while, by searching for distant predecessors or successors of fields. This feature is absent from many papers whose methods consider only contiguous periods for inter-temporal matching.

- **Agnostic about the different dimensions of meaning** (y/n). Saussure's theory of language (, p123) distinguishes two kinds of relations between terms in a sentence that convey different classes of values and correspond to two forms of our mental activity. "In discourse, on the one hand, words acquire relations based on the linear nature of language because they are chained together. [...] Combinations supported by linearity are *syntagms*. [...] Outside discourse, one the other hand, words acquire relations of a different kind. Those that have something in common are associated in the memory, resulting in groups marked by diverse relations. [...] The syntagmatic relation is *in praesentia*. It is based on two or more terms that occur in an effective series. Against this, the

associative relation unites terms in *absentia*, in a potential mnemonic series." The associative relation between terms is also called the *paradigmatic* axis. While co-word methods can chose to adopt similarity measures that focus on either the syntagmatic axis or the paradigmatic axis, methods that rely on word embedding focus only on the paradigmatic axis (structural equivalence between words). As for methods from topic modeling and citation analysis, the distinction between these two axes of language is not made explicit in the methodology and the description of the subjects mixes these two dimensions in an indeterminate manner.

- **Allow split/merge events** (y/n): In the evolution of science, certain disciplines regularly give rise to several and others hybridize to form new ones (e.g. computational social sciences, bio-informatics, econopysics, etc.). These important events should be described by splits and mergers in the phenomenological reconstruction of science. While some methods take these events into account, others consider only linear inter-temporal matching.
- **Multi-level** (y/n): Some methods take into account the multi-level structure of science, as described in 1.2, others do not even mention this concept,
- **Multi-scale** (y/n): Some methods allow to manage the complexity of the reconstruction by filtering information or aggregating certain topics to simplify the representations. Although a significant proportion of the papers do not conceptualize this as a multi-scale exploration of the phenomenological reconstruction, varying certain parameters of the reconstruction in fact often allows to generate reconstructions at different scales of observation.
- **Objective function** (y/n): Some methods are based on an objective function that is optimized by tuning the parameters of the reconstruction. The presence of such an objective function gives some interpretation keys to the reconstruction and allows the continuous improvement of the methods with respect to this objective function through the evaluation of alternatives for certain parts of the workflow.
- **Internal/quantitative evaluation** (y/n): Some papers compare their approach with previous published work or alternative workflows. Others do not evaluate their methods up do this point.
- **External/qualitative validation** (y/n): Some papers make an external validation of the proposed methods by applying their method to some domain of science and then comparing their results to some ground-truth like the known history of a that domain, independent expert knowledge or literature reviews. Others do not evaluate their methods up do this point.
- **Integrate users' preference** (y/n): Phenomenological reconstruction makes it possible to design digital tools that provide to an individual or a collective with an additional understanding of their interactions and actions on their surrounding world. These tools should be adapted to their users' preferences and goals. Some of previously published work take into account this dimension (often by way of interactivity), others don't.
- **Visualization** (y/n): Phenomenological reconstruction of science evolution can be analyzed by way of statistics or descriptions of their parts but *ad-hoc* visualization is required to grasp their totality. Some papers develop some *ad-hoc* visualization, other do not go up to this point.
- **Software reproducibility** (y/n): Reconstruction of science dynamics requires complex computer processing. Papers differ in the level of reproducibility of the proposed method, the highest level of reproducibility being when the method is implemented into a usable software.
- **Open Source scripts** (y/n): Whether implemented as a simple script or within software, the computer code of a method must be open to ensure full reproducibility and verifiability of the proposed results.

Tables 2 and 1 compare key papers on knowledge and science dynamics reconstruction with regard to these criteria. This paper is the one that has the highest coverage of all these issues.

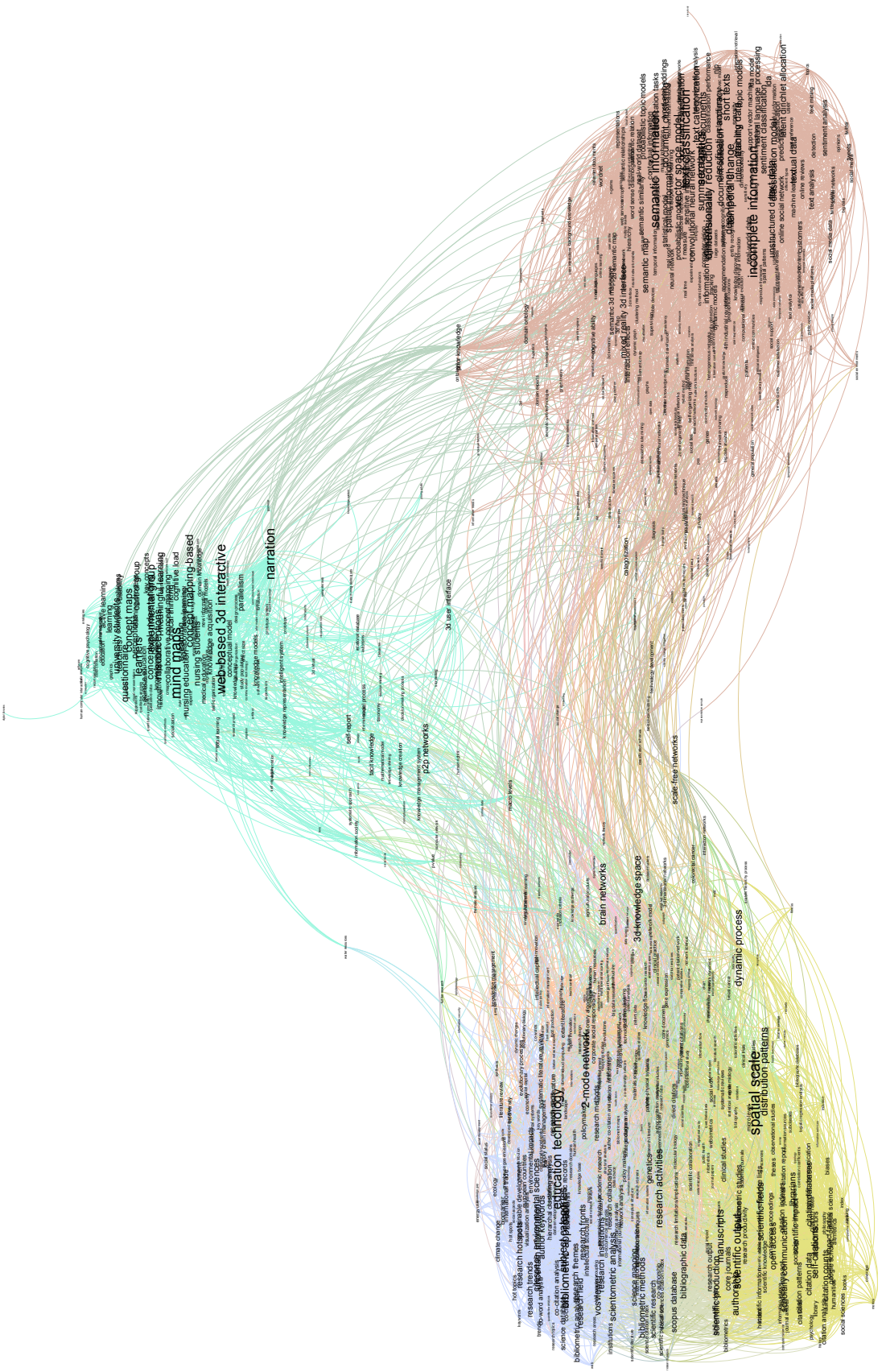

Figure 12: Map of the domain of *science and knowledge mapping* ($\mathcal{D}_{maps}$: 13,844 publication meta-data extracted from the Web of Science by using research equation B.1). Links are computed according to a distributional similarity metric (second order or paradigmatic similarity). Generated with Gargantext and spatialized in Gephi with the Force Atlas algorithm . An interactive version of this map is available at `http://maps.gargantext.org/maps/sciencemaps` and a stand alone version can be downloaded from the Harvard Dataverse https://doi.org/10.7910/DVN/SBH3EI.

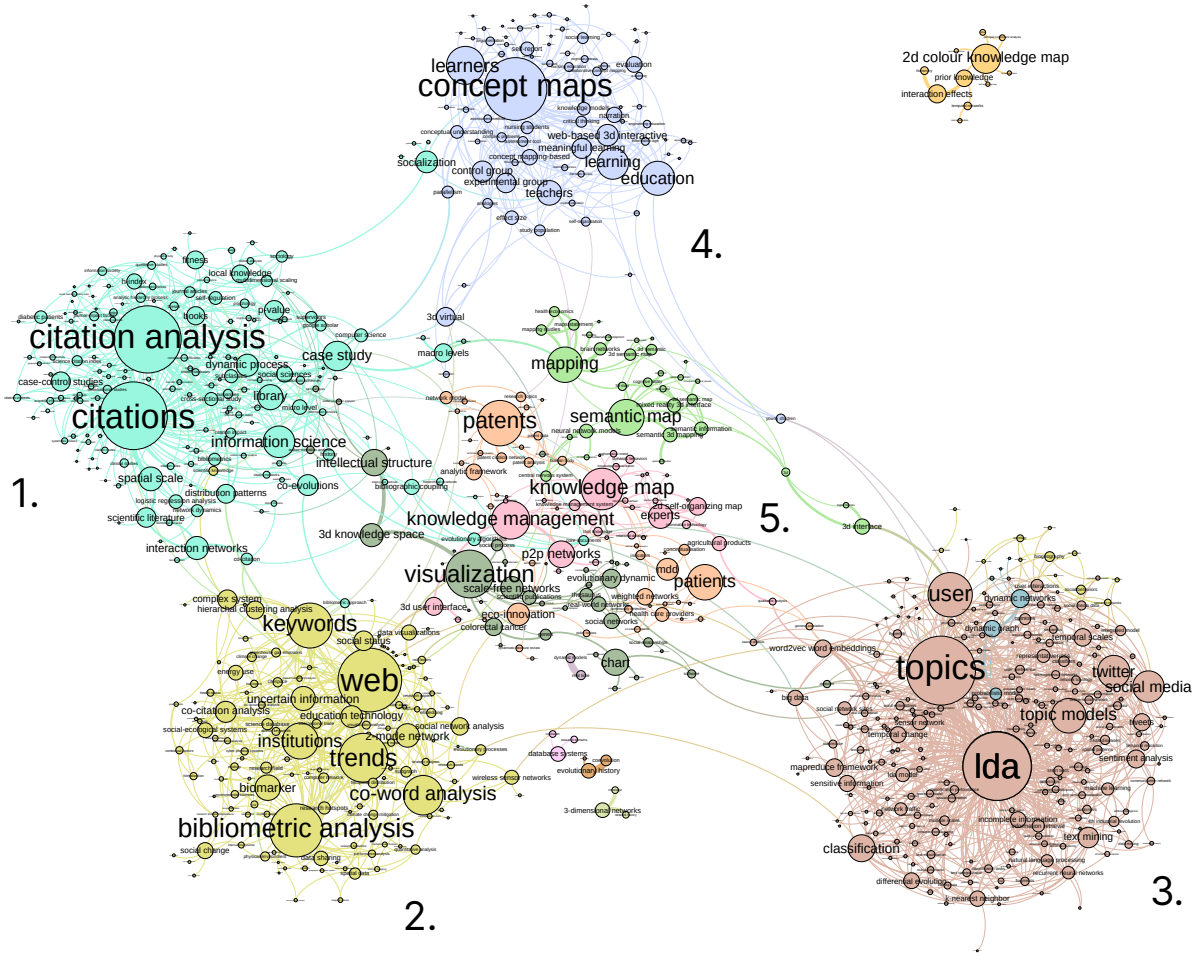

Figure 13: Map of the domain of *science and knowledge mapping* ($\mathcal{D}_{maps}$: 13,844 publication meta-data extracted from the Web of Science by using research equation B.1). Links are computed according to the confidence similarity metric (first order or syntagmatic similarity measure). Generated with Gargantext and spatialized in Gephi with the Force Atlas algorithm . An interactive version of this map is available at `http://maps.gargantext.org/maps/sciencemaps` and a stand alone version can be downloaded from the Harvard Dataverse https://doi.org/10.7910/DVN/SBH3EI.

C Glyphosate history

In order to evaluate our method in relation to an external referential, we compare in this section a phylomemy based on the literature of a scientific field to the history of the same field as told by its protagonists. We chose for this benchmark the literature on glyphosate that has several remarkable characteristics.

Glyphosate has been introduced as a pesticide in the mid seventies and became the most marketed herbicide active ingredient by the nineties, and more or less holds that position ever since . The literature on glyphosate is therefore quite recent, most of it being available in digital format, and the knowledge it contains is of great importance from a health and economic point of view.

Moreover, glyphosate literature has the peculiarity that there is no consensual synthesis of the knowledge it contains. Glyphosate is a controversial herbicide about which literature reviews and historical analyses are regularly published, some emphasizing advantages of glyphosate (e.g.) or the absence of associated risks (), others reviewing the risks for health and environment and issues related the the emergence of herbicide resistant weeds . Although most historical studies recognize weeds glyphosate resistance as a major issue , most reports about other possible glyphosate negative effects are made controversial. For example, when in 2015 the International Agency for Research on Cancer (IARC) published a monograph concluding that glyphosate "is probably carcinogenic to humans" , the assessment has been followed by an immediate backlash from industry groups that has produced several papers pointing to bias in the IARC study . "Additionally, in 2015, just a few months after the IARC monography published on glyphosate, the EFSA (European Food Safety Authority, Parma, Italy), another WHO related organization, declared that it was "unlikely" that the molecule could be carcinogenic to humans or that it could cause any type of risk to human health. The conflict between the two organizations of the World Health Organization triggered many doubts, and for this reason, a series of independent studies were launched to better understand what glyphosate's danger to humans and the environment really was" .

In such scientific context, there is a high risk of selection bias in the choice of the publications considered in literature reviews and monographs, that is, there is a risk of cherry-picking the publications that confirm theoretical claims rather than those that do not. This is precisely the criticism made by glyphosate advocates to the report of the IARC. In fact, no-one can claim to have read all glyphosate related publications and the references mentioned in a literature review are necessarily a selection that reflects what is judged the most important from the point of view of its authors. In such situation, as highlighted by , phylomemy reconstruction could be useful to give an overall picture of the field which would be as objective as possible in the sense that it would consider as many publications on the topic as possible and process them equally, only on the basis of a definition of what constitutes a valid publication. Such phylomemy could highlight, before any further considerations of what is important and what is not, the main issues addressed by the scientific community and their global trends.

In this appendix, we compare a phylomemy built from an interdisciplinary corpora $\mathcal{D}_{glyphosate}$ on glyphosate constituted of 16,655 documents from 1990 to 2020[24] (*cf.* Figure 14) with a synthesis of key historical reviews written by proponents and skeptics of glyphosate use .

C.1 Main research topics on glyphosate use

The main subdomains of glyphosate literature as referenced by and concern the biochemistry of glyphosate, weed management with pre-emergent and later post-emergent application technology, the environmental fate of glyphosate and adverse environmental effects of glyphosate. For more completeness of the external identification of glyphosate sub-branches we summarize in table C.1 the main section titles of three major literature reviews on glyphosate from different classes of protagonists of this domain of research. As a whole, they define the main research areas about glyphosate.

[24] Timestamped titles and abstracts extracted from the Web of Science on June 2020 with the query "glyphosate OR roundup OR 'round-up"'. Because of the lack of abstracts in digital archives before 1990, we have decided to restrict our comparison on the period 1990-2020 to avoid artifacts due to the database coverage.

Table B.1. Summary of main research areas related to glyphosate from the main sections of literature reviews , and .

Singh et al. 2020 *Environ Chem Lett.*	Duke 2018. *Pest Manag. Sci.*	Székács & Darvas 2012 *Herbicides-Properties, Synthesis and Control of Weeds*
• Uptake and translocation of glyphosate in plants		• **The discovery of glyphosate**
		• Mode of action
• Mechanistic action of glyphosate in plants	• Pre glyphosate-resistant crops	• Transition state analogue theory of enzyme inhibition
		• Other biochemical effects of glyphosate
• The emergence of glyphosate-resistant (GR) crops		• **Pre-emergent application technology of glyphosate**
		• Formulated glyphosate-based herbicides
• The mechanism involved in the evolution of GR crops	• The golden age of weed management; the first ten years of glyphosate-resistant crops	• Formulating agents
		• **Post-emergent application technology of glyphosate**
• Toxicity of glyphosate		• Glyphosate-tolerant crops
		• The effect of glyphosate-tolerant crops on glyphosate residues
• Effect on amphibians and fishes		
	• Detractors	• **The environmental fate of glyphosate**
• Effect on higher vertebrates		• Residue analysis of glyphosate
		• Glyphosate and its decomposition products
• Effect on humans		
		• Environmental monitoring of glyphosate
	• Darwinian reality spoils a good thing	• **Adverse environmental effects of glyphosate**
• Analytical detection and quantification of glyphosate		
		• Glyphosate and Fusarium species
• Chemical degradation of glyphosate		• Toxicity of glyphosate to aquatic ecosystems and amphibians
	• Unexpected phenomena of glyphosate	
• Microbial degradation of glyphosate		• Teratogenic activity of glyphosate
		• Genotoxicity of glyphosate
• Complexation chemistry of glyphosate with metal ions and humic acid		• Hormone modulant effects of glyphosate and POEA
		• Glyphosate resistance of weeds

C.2 Exploration of glyphosate phylomemetic networks

First, let us note that from the mere observation of the absolute or relative publication volume of glyphosate 14, it can only be said that interest for glyphosate has increased over time, with apparently two periods of acceleration starting in the mid 1990s and in the mid 2010.

The analysis of the shape and content of the phylomemy reconstructed for a sensibility value of 0.7 at different scales (Figure 15 to Figure 19 for scales 0.99, 0.9, 0.85, 0.8 and 0.5 respectively) explain these two accelerations and highlights other important phases of development of glyphosate research in the early 2000s and in the early 2010s. It put forward a shift in glyphosate literature from research on its effects on plants, its benefits for agriculture and its usages (1990-2005, a period called the 'golden age' of glyphosate, Figure 20) to the problem of herbicide resistant weeds management (2005 - , Figure 21), the questions of risks for the environment, and health issues for animals (2004 -) and its threats to humans health (2014 - Figure 22).

The detailed screening of the different phylomemy branches with the help of a dedicated visualization software reveals that the phylomemy reconstruction also successfully maps the details of the different issues and areas of research covered by the glyphosate literature present in table C.1. This can be verified in the interactive explorer of the glyphosate phylomemy available at https://doi.org/10.7910/DVN/SBH3E.

Consequently, phylomemy reconstruction does not only succeeds in mapping the different issues and areas of research covered by the glyphosate literature, but also successfully reconstructs the timeline for the development of the issues identified by the main literature review of the domain.

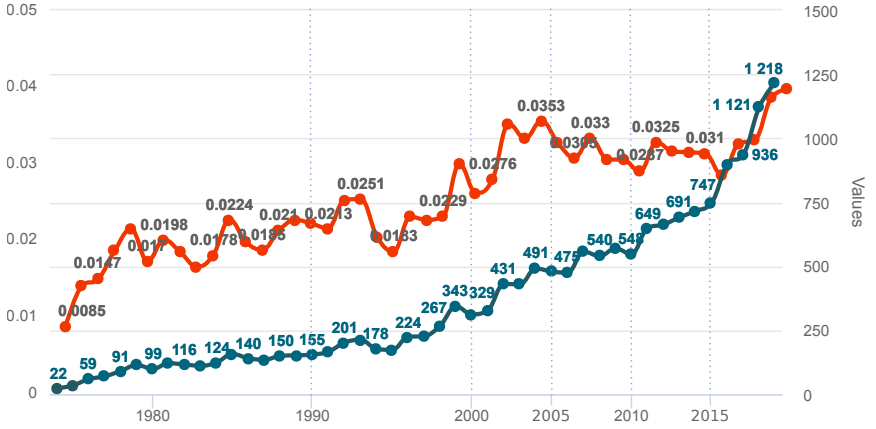

Figure 14: Evolution of the number of publications retrieved with the query *"glyphosate OR 'round-up' OR roundup"*, 16,655 publication between 1975 and 2020. Blue: absolute numbers, red: relative proportions in the Web of Science publications.

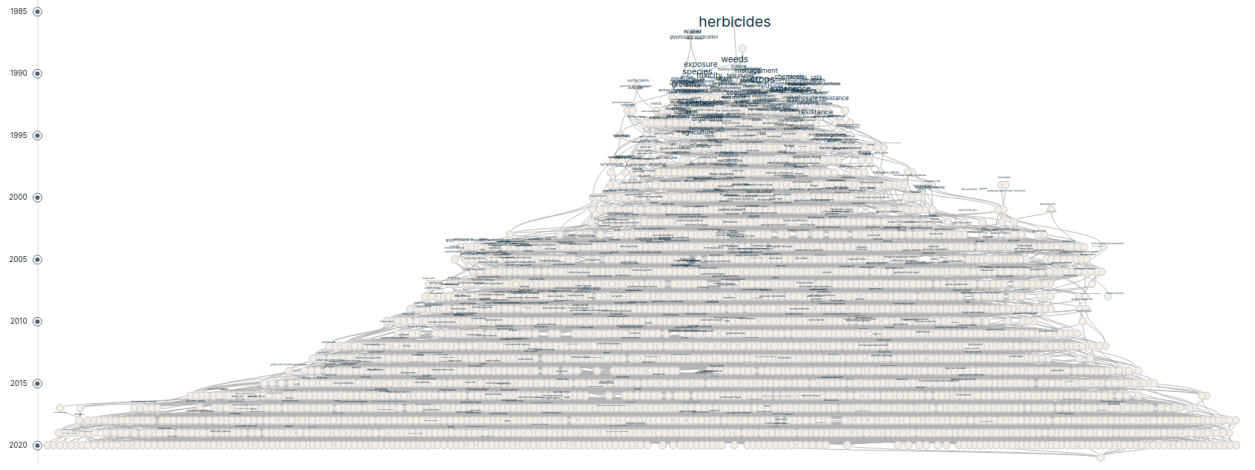

Figure 15: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ at the level $\lambda = 0.99$ with branches smaller than 3 filtered out. 1,152 terms (from a list of 1,322 initial terms) appear in 3,110 groups of this single branch phylomemy computed over 16,655 documents.

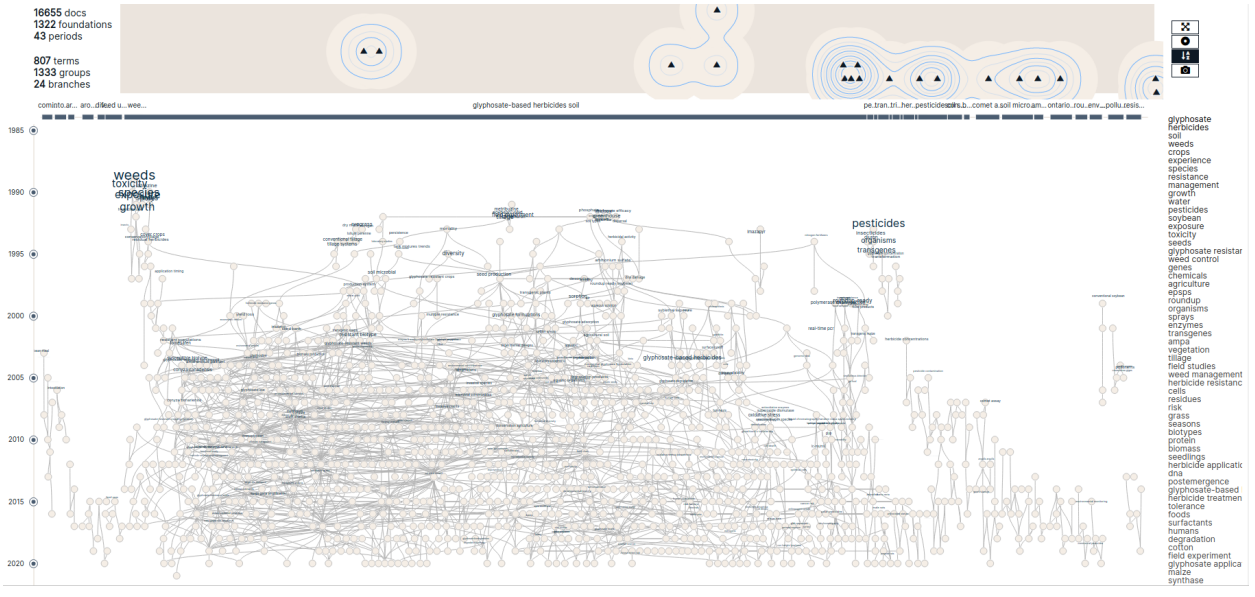

Figure 16: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ at the level $\lambda = 0.9$ with branches smaller than 3 filtered out. 807 terms (from a list of 1,322 initial terms) appear in 1,333 groups of this phylomemy computed over 16,655 documents. The list on the left of the figure display the most frequent terms from the scientific fields of the main branch, ordered by the number of fields that mention them.

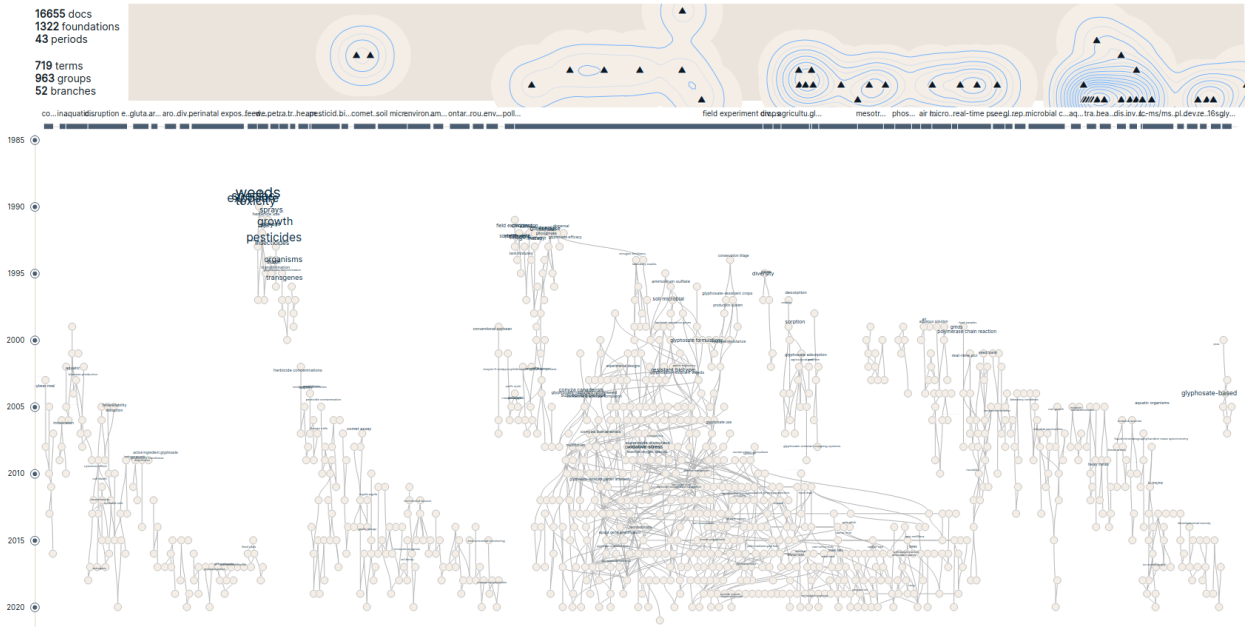

Figure 17: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ at the level $\lambda = 0.85$

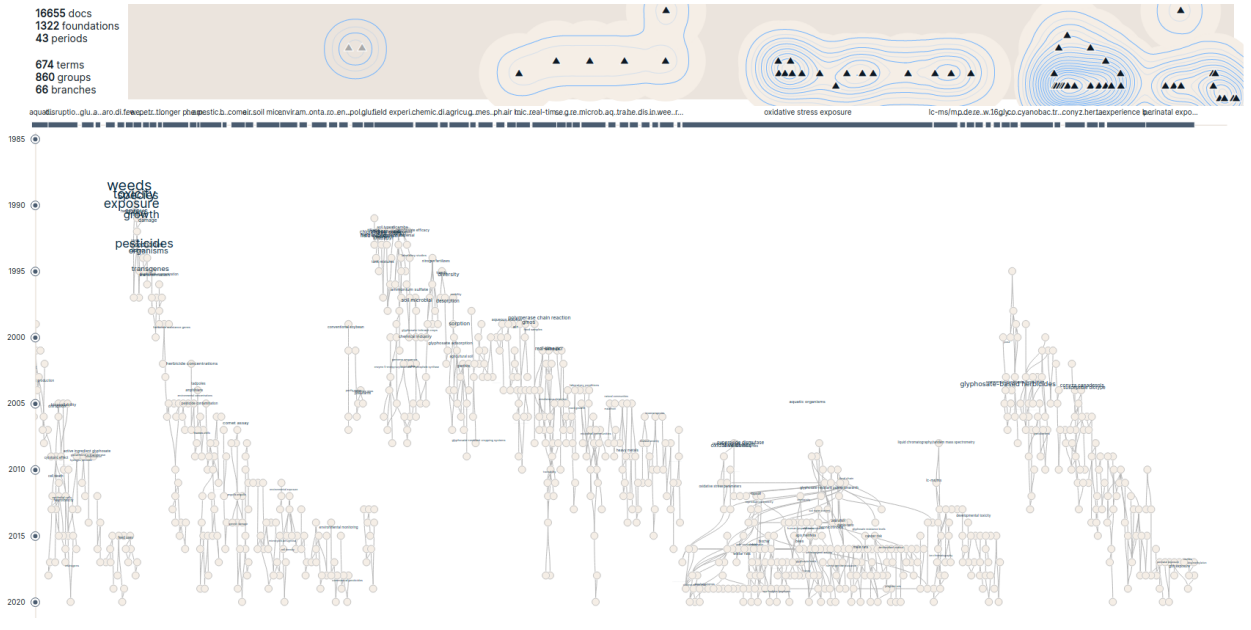

Figure 18: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ at the level $\lambda = 0.8$. An interactive version of this phylomemetic network is available at `http://unpublished.iscpif.org/glyphosate` and can be downloaded from the archives (data), (explorer).

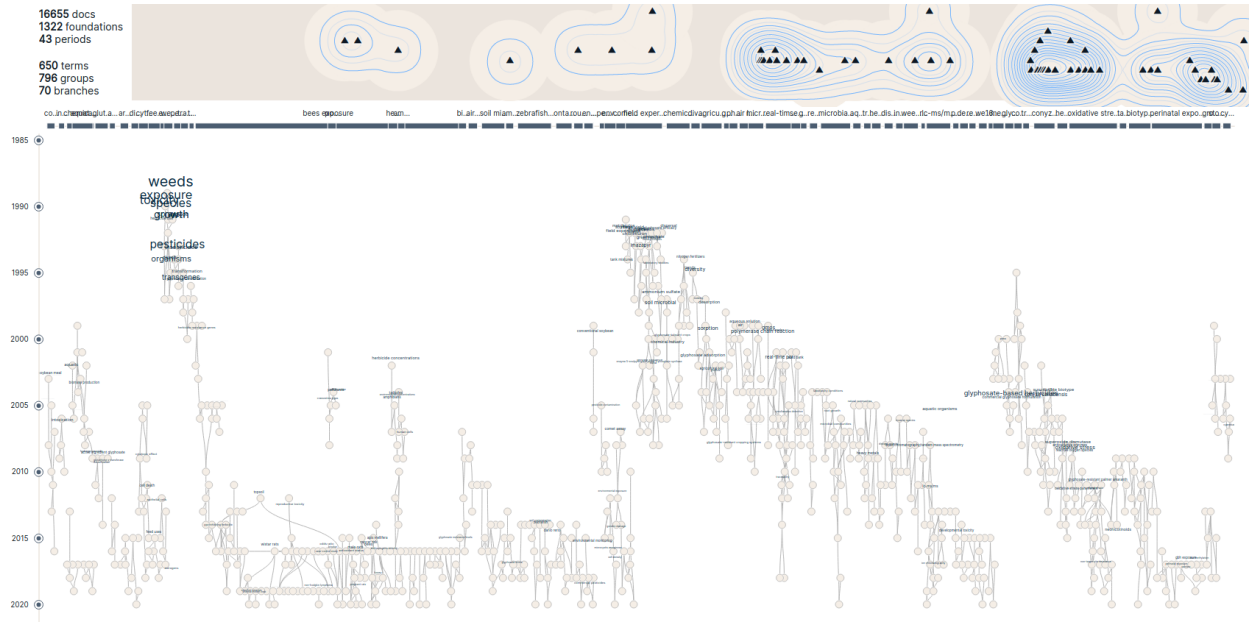

Figure 19: The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ at the level $\lambda = 0.5$

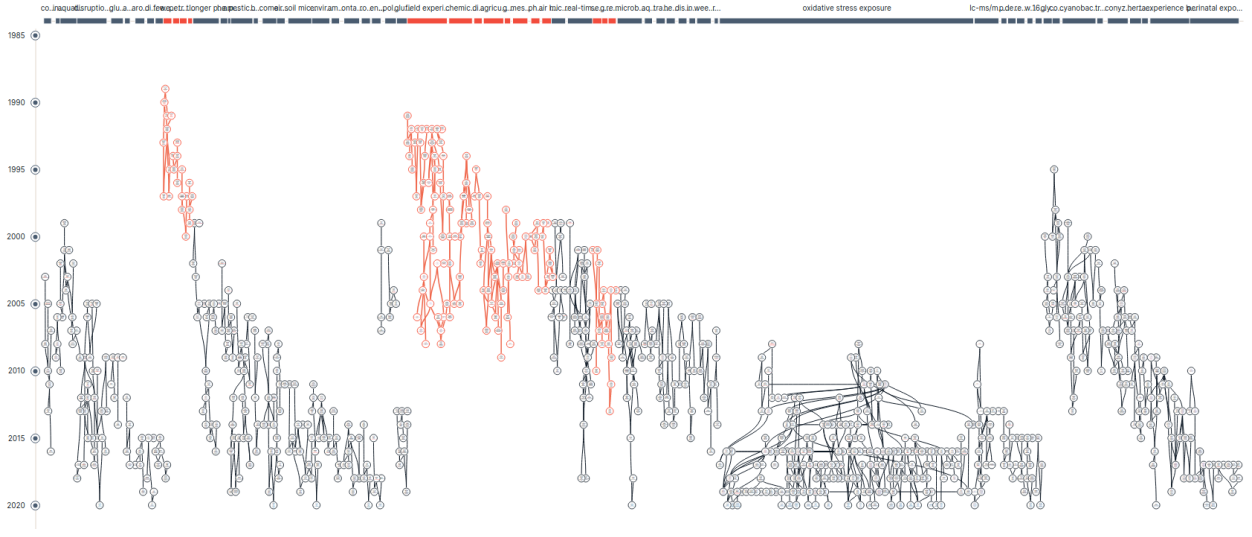

Figure 20: The golden age of glyphosate. Research on benefits of glyphosate and its post-emergence applications (1990-2010) $\lambda = 0.8$. The corresponding branches are highlighted in red.

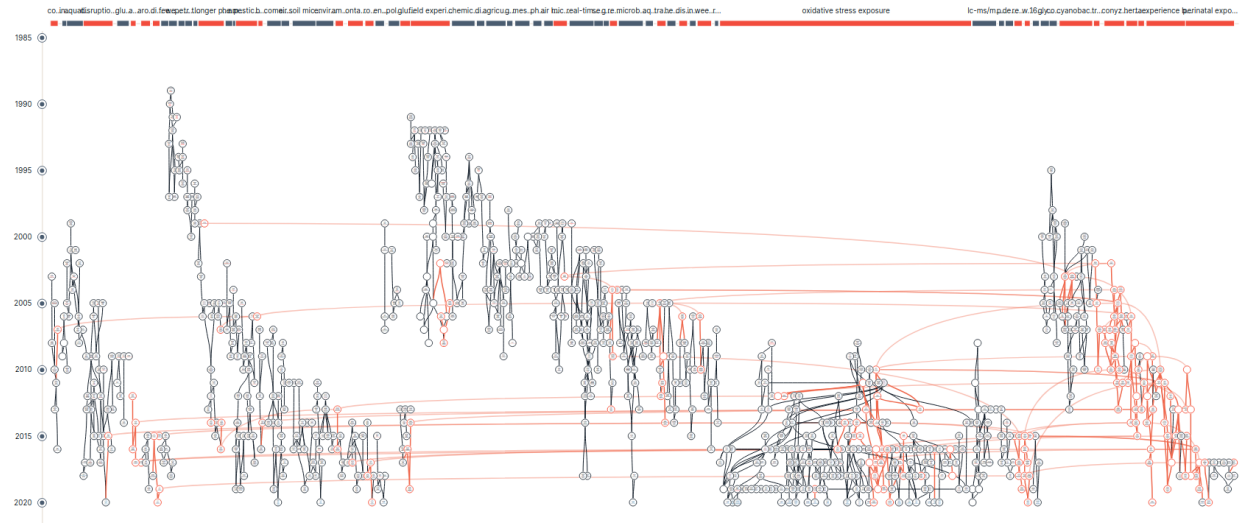

Figure 21: Research on the emergence of glyphosate resistance and its mechanisms ($\lambda = 0.8$). As can be seen on this figure, this issue was mostly address over the period 2005-2020. Groups highlighted in red mention one or several of terms related to the issue of glyphosate resistance : resistance / selection pressure / mutation / multiple resistance / herbicide resistance / epsps gene amplification / arabidopsis thaliana / lolium perenne / lolium multiflorum / eleusine indica / conyza canadensis / amaranthus palmeri / ambrosia trifida / kochia scoparia / glyphosate-resistant horseweed / echinochloa colona / conyza bonariensis / sorghum halepense / ryegrass / herbicide-resistant weed / resistant biotype. Several branch deal with this issue, Link between branches highlight the reuse of concept from one branch in an other.

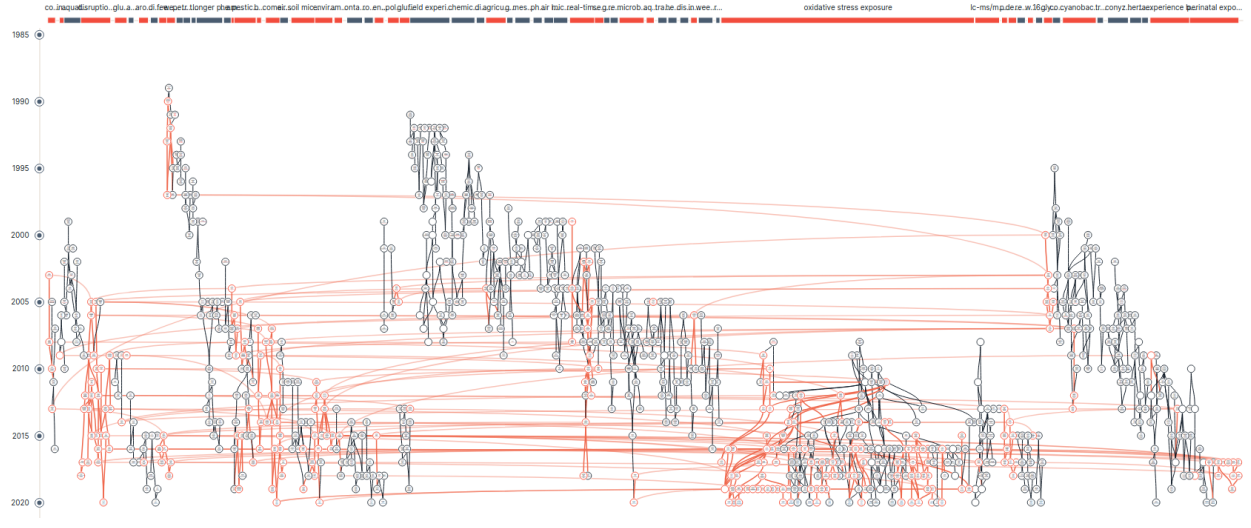

Figure 22: Research on the emergence of glyphosate-related health and environmental issues $\lambda = 0.8$ As can been seen on this figure, this issue intensified in the 2010s to become a major concern nowadays. Groups highlighted in red mention one or several of those terms : health, toxicity, acute toxicity, human, animals, rats, cells, genotoxicity, dna damage, hormesis, low doses, estogens hormerosis, aquatic organisms, contamination, toxicol, harmful effect, poisoning, risk, food safety.

C.3 Clinical trials

Several hundred clinical trials on COVID-19 were launched within a few weeks worldwide. Planned/ongoing/terminated trials are being tracked by WHO and physicians.

The project Covid-NMA coordinated by the CRESS (CNRS-UMR1153) is collecting the WHO clinical trials records and pre-process them in order to filter, clean and enrich the dataset. The aim of this database is to inform about current studies and explore new therapeutic targets, avoiding similar studies. The database analyzed here has been generated on 2021/01/20. It synthetizes more than 5994 arms of clinical trials related to COVID-19.

The provision of a synthetic view of these clinical trials is essential for the coordination of the research community: to explore new therapeutic targets while avoiding similar studies. This can be achieved thanks to phylomemy reconstruction, based on textual data mining of the metadata of these trials.

We have considered all the treatments mentioned in this data base as our list of roots $\mathcal{L}$ and applied the phylomemy reconstruction workflow to the corpora of clinical trials without any filter on clique sizes. Two treatments are in the same cluster of this phylomemy at a given period if they have been jointly tested for curing covid-19 patients.

The resulting phylomemetic network (cf. Figure 23) clearly and exhaustively depicts all the research path explored so far in the hope of finding a cure to COVID-19. Among others, we can observe that clinical trials on hydroxychloroquine belong to the most complex branch and that this compound has been combined with a large number of other treatments (i.e. azithromycin, ritonavir, lopinavir, zinc, oseltamivir, interferon beta, umifenovir, darunavir, tocilizumab, naproxen, dalargin, vitamin d3/b12, raltegravi) contrary to the vast majority of other treatments that form linear branches that reflect an absence of research on potential treatment combination.

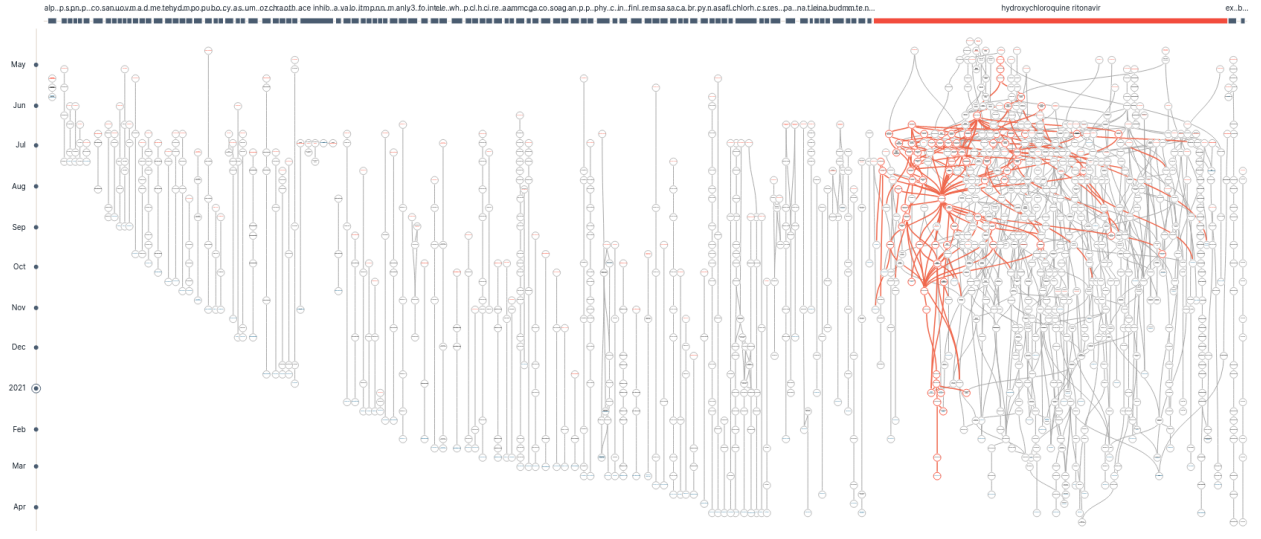

Figure 23: Phylomemetic network of the treatments mentionned in the WHO clinical trial database generated at a level of observation $\lambda = 1$ without any filter on the cliques. Fields highlighted in red are those mentioning hydroxychloroquine. As observed by biomed scientists, hydroxychloroquine has mobilized huge investigation efforts between June and October 2021, being tested in combination with several other treatments before being discarded : azithromycin, ritonavir, lopinavir, favipiravir, darunavir, nitazoxanide, zinc, ivermectinoseltamivir, sofosbuvir, daclatasvir, bromhexine, methylprednisolone, interferon betaX, dexamethasone, sarilumab, tocilizumab, camostat mesilate, baricitinibumifenovir, cobicistat, thalidomide, immunoglobulin, convalescent plasma, treatment ribavirin, ciclesonide, imatinib, sirolimus, ciclosporin, prednisone, canakinumab, famotidine, omega-3, tofacitinib, lithium, trifluoperazine, atazanavir. An explorable version of this phylomemetic network is available at https://doi.org/10.7910/DVN/WLI9B5.

D Detail and implementation of the phylomemy reconstruction workflow

In what follows, we will detail the four operators of the phylomemy reconstruction process $\Phi = {}^4\Phi \circ {}^3\Phi \circ {}^2\Phi \circ {}^1\Phi$ by first providing additional elements of vocabulary according to their order of appearance in the workflow. For the sake of clarity, we will refer to the formalism and notations previously introduced in section 3.1 and illustrated by the figure 1. To that end, we will also take the example of science and of its digitized archives as a particular type of collective knowledge, although our methodology can be applied on top of any kind of elements of knowledge, from Web pages to industrial patents. Finally and for reproducibility purpose, we will review the Haskell implementation[25] of the whole workflow within the free software *Gargantext*.

D.1 Indexation (${}^1\Phi$)

All co-word analyses start by combining a corpus and a structured list of terms (called *map list*). Both of them are bound to evolve in a recursive and reflexive way during the phylomemy reconstruction process: the corpus is cleaned and completed, terms are added, merged or suppressed, etc. The choice of a particular pair of corpus and map list, associated with the choice of some reconstruction parameters (expressed as a Json file by the user), define uniquely a phylomemy. With that in mind, the operator of indexation ${}^1\Phi$ can be summarized as follows:

- ${}^1\Phi.1$ Identification a corpus that records a particular type of knowledge,
- ${}^1\Phi.2$ Identification of the core vocabulary that describes this knowledge in the form of a list $\mathcal{L} = \{r_i \,|i \in \mathcal{I}\}$ of groups r_i of terms (called *roots*) that convey the same meaning from the point of view of the researcher,
- ${}^1\Phi.3$ Choice of time slicing method. This defines the temporal resolution of the final output (e.g., 3 year) and an ordered series $\mathcal{T}^* = \{T_i\}_{1 \leq i \leq K}$, $T_i \subset \mathcal{T}$, of sub-period of $\mathcal{T}^*$,
- ${}^1\Phi.4$ Choice of the type of *contextual units* for computing co-occurrences within documents (e.g., full documents, paragraphs, sentences, etc.). This gives a resolution to the semantic proximity we will later compute,
- ${}^1\Phi.5$ Co-occurrence count of elements of vocabulary within the chosen contextual units of documents per period of time,

Corpus (${}^1\Phi.1$) A phylomemy is shaped on top of large corpora of scientific contributions extracted from academic databases. The choice of the corpora is part of the research question of the investigator and its completion could be one of the output of a phylomemy reconstruction that would highlight missing parts of literature. The identification of the scope of the timestamped literature to be analyzed is consequently a reflexive process enhanced by phylomemy reconstruction. It has however to start somewhere and is usually defined by a query in a search engine. We don't investigate this precise part here and assume that the researcher has chosen to investigate a particular corpora which size is usually comprised between several thousands to few millions documents.

Core vocabulary (${}^1\Phi.2$) The map list defines the vocabulary of interest for the domain under study. It can be established by different means: from a scientific ontology like the *Medline Mesh*, from some domain specific dictionaries or it could be extracted from the corpus itself by applying natural language processing. The most elementary part of a vocabulary is an *n-gram*, that is, a contiguous sequence of n words from a given sample of text or speech ($n \geq 1$). In a quantitative analysis, there are many situations in which two n-grams convey exactly the same meaning. For example, we rarely care about spelling differences between U.S. English and U.K English nor between synonyms, between plurals and singular, between a n-grams an its acronym or about typos. But from a statistical point of view and in order to make the upcoming visualization more readable, it is important to treat some of these terms as single entities. That is why, we start by grouping some terms to set up our targeted level of analysis. To that end, we can here make use of linguistic equivalences (like synonyms, singulars and plurals, translated expressions, etc.) or we can tune the grouping operation by ourselves according to our own needs within *Gargantext*. A *root* is thus defined as a set of n-grams (e.g., {*concept mapping tools*, *concept mapping tool*, *concept map tool*, *concept map tools*}) that, from the point of view of the investigator, convey a meaning equivalent to a given label (e.g, *"concept mapping tools"*). In the phylomemy reconstruction workflow, co-word statistics are computed at the level of roots while visualizations only display the

[25] Full Haskell code is available at `https://gitlab.iscpif.fr/gargantext/haskell-gargantext/tree/master/src/Gargantext/Core/Viz/Phylo`

labels of a root item. The set of roots forms a list of pairs of type $[n-gram, \{n-gram\ set\}]$ that shape the map list noted $\mathcal{L} = \{r_i = [l_i, \{x_k\}] | i \in \mathcal{I}\}$ where l_i is the term used to label the root r_i in visualizations and $\{x_k\}$ is the set of terms that are considered as equivalent to l_i for the investigator. In the following, we will always place ourselves at the level of roots and we will mention interchangeably the roots or one of the terms of which they are composed if this makes it easier to understand. The whole set of roots are called *foundations* of the phylomemy.

Temporal resolution ($^1\Phi.3$) As stated in the introduction, we aim for the reconstruction of knowledge evolution, which means that we will follow the dynamics of terms interactions over several sub-periods of time. We expect this evolution to be multi-scale with changes of various magnitudes from n-grams meanings to interactions patterns that will depend on the length of the sub-periods. Since co-words analysis is a quantitative methodology based on statistical effects, there must be enough documents to process over each sub-period in order for the similarity measures to be statistically significant. The minimal time resolution has thus to be chosen according to the acceptable tolerance over the imprecision of similarity measures. It can be measured for example through the distribution of confidence intervals on pairwise similarity measures. Once defined a minimal temporal resolution, the investigator can chose a specific temporal resolution for its observation of the temporal dynamics which results in a slicing of $\mathcal{T}$ into periods of time $\mathcal{T}^* = \{T_i\}_{1 \leq i \leq K}$, each T_i being larger than the minimal temporal resolution. $\mathcal{T}^*$ could be heterogeneous depending on the nature of the corpora. For example, Rule *et al.* have proposed an original slicing method based on homomorphic reduction of the transition matrix capturing change in key terms used over time. This makes it possible to adapt temporal co-word analysis to corpora for which major changes in the vocabulary of documents occur at irregular moment in time [26]. As for Science, given the fact that the many subdomains have their own dynamics with asynchronous changes in the pace of discoveries, when addressing a *a priori* multi-disciplinary corpora it is more appropriate to consider a regular and overlapping slicing of $\mathcal{T}$. For a first overview of science dynamics, a resolution of 3 to 5 years, that corresponds to the characteristic time for a PhD completion is a good start when the domains under investigation are producing enough documents per year. Adopting a temporal sliding window with time intervals that are shifted by one year makes it possible to observe the delta that each new year bring to the body of knowledge.

Contextual units and Co-occurrences ($^1\Phi.4$,$^1\Phi.5$) After having chosen the scope of our contextual units[27], the indexation $^1\Phi$ ends by processing the co-occurrences between the roots. This computation is ideally made at the smaller possible temporal resolution allowed by the corpus (all academic papers have a year of publication, all tweets are dated down to the second, etc.). If, for example, the corpus is an academic archive covering a period $\mathcal{T} = [1990 - 2020]$, the result of the indexation step will be a set of co-occurrence matrices $\mathcal{M} = \{M_t, t \in \mathcal{T}\}$ where each cell $m^t_{i,j}$ of M_t will represent the number of contextual units mentioning both root i and root j at year t.

Implementation of $^1\Phi$ in Gargantext The operator of indexation $^1\Phi$ takes two input files: 1) the original corpus of documents and 2) the maplist. Both files follow the CSV's format of Gargantext, such as: 1) `title; source; publication_year; publication_month; publication_day; abstract; authors` and 2) `status; label; forms`. Shaped by the researcher, the maplist here associates lists of equivalent semantic forms to chosen labels. Here, we first make use of Gargantext's text mining mechanisms to tokenize and lemmatize each document. We then match the resulting n-grams with their corresponding forms in the maplist and replace each of them with their respective label. At this stage, the documents are reduced into timestamped lists of root terms. In the same time, we define a sorted list of fixed periods on the basis of two parameters of time *range* and *step* (both in years). A given period might thus partially overlay its successor, and so on. We next distribute the documents among the periods by means of their publication year. We finally initialize a basic phylomemy as a data structure[28] containing periods and n-grams, along with useful elements of information like the raw number of documents per periods or the n-gram frequency throughout time. Within each period, we turn the lists of n-grams into a co-occurrence matrix by counting the number of times two terms jointly appear in a shared document.

D.2 Similarity measures ($^2\Phi$)

After having obtained the set of co-occurrence matrices $\mathcal{M}$, we move to the second step $^2\Phi$ of the workflow that defines what is considered to be meaningful relation between root terms. In the fields of linguistics, scientometrics or

[26] They considered more than 200 years of *The annual State of the Union addresses* which tone where been impacted by major U.S politics events

[27] Thereafter, we will consider a document with its title and body as a contextual unit.

[28] See the *Phylo* Haskell type in https://gitlab.iscpif.fr/gargantext/haskell-gargantext/blob/master/src/Gargantext/Core/Viz/AdaptativePhylo.hs

computer sciences, several similarity measures have already been defined over a set of n-grams. To guide our choice of similarity measures out of co-occurrence statistics, we adopt the framework of Saussure's theory of language (, p123):

"Relations and differences between linguistic terms fall into two distinct groups, each of which generate a certain class of values. [...] They correspond to two forms of our mental activity, both indispensable to the life of language. In discourse, on the one hand, words acquire relations based on the linear nature of language because they are chained together. [...] Combinations supported by linearity are syntagms. [...] In the syntagm a term acquires its value only because it stands in opposition to everything that precedes of follows it, or both. Outside discourse, one the other hand, words acquire relations of a different kind. Those that have something in common are associated in the memory, resulting in groups marked by diverse relations. [...] The syntagmatic relation is in praesentia. It is based on two or more terms that occur in an effective series. Against this, the associative relation unites terms in absentia, in a potential mnemonic series."

The review of those measures is out of the scope of this paper. But yet, we can here broadly detail two categories of such metrics, that is 1) the first order metrics and 2) the second order metrics, also called 1) the *syntagmatic* axis and 2) the *paradigmatic* axis in linguistics:

1. First order / syntagmatic axis. A first phenomena of interest is the direct interaction between n-grams in documents. In this case, the similarity measure $P(x, y)$ between n-grams x and y is a function of the number of documents that mention both of them. An example of such similarity measure if the *confidence* defined as the maximum of the two probabilities of having a term knowing the presence of the other in the same contextual unit ($P(x, y) = max(p(x|y), p(y|x))$). It has been demonstrated for example that *confidence* is one of the most appropriate metrics to recover the generic/specific relationships between n-grams in an ontology-like way .

2. Second order / paradigmatic axis. A second interesting aspect of n-grams similarities is their structural equivalence, that is the similarity of meaning between fragments of text. Most of the related works there assume the Harris hypothesis and derive the similarity between two n-grams x and y from the similitude of their contexts of appearance in the corpus. Each n-gram is thus represented by a vector or a matrix that describes its various contexts of appearance . Reductions of dimensions or/and neural networks are then used to end with a similarity measure between x and y that would be proportional to the similarity of their contexts of appearance in the corpus.

It should be clearly stated that choosing a similarity measure will define a particular class of meanings on n-grams interactions as well as as singular operators of projection (in the sense described in 2.3). In other words, different similarity measures will produce distinct phylomemies. So once the similarity measure and the temporal slicing $\mathcal{T}^*$ are chosen, the set of timestamped similarity matrices can be processed. The output of ${}^2\Phi$ is thus a set of graphs $\mathcal{G} = \{G_T = (V, E_T) | T \in \mathcal{T}^*\}$ where V is a set of vertices representing the roots of the phylomemy and E_T is a set of edges measuring the similarity between those roots over the period T. The matrices describing the weighted edges are noted $\mathcal{W} = \{W_T, T \in \mathcal{T}^*\}$ where each cell $w^T_{x,y}$ of W_T represents the similarity between x and y (i.e., weight of the edge $x \overset{w^T_{x,y}}{\longleftrightarrow} y \in E_T$).

Implementation of ${}^2\Phi$ in Gargantext The co-occurrence matrices $\mathcal{M}$ here become similarity matrices $\mathcal{W}$ by calculating the *confidence* measure between the n-grams within each period.

D.3 Field detection (${}^3\Phi$)

The matrices $\mathcal{G}$ and $\mathcal{W}$ synthesize one quantitative point of view on our corpora. At that stage, we can apply methods from graph theory and complex networks analysis to detect remarkable structures inside $\mathcal{G}$ and $\mathcal{W}$. These methods are named clustering algorithms or community detection algorithms depending on the scientific community that develop them. Community detection in graphs is a field of research on its own and, once again, several options are available to us. Following , we will define thereafter scientific fields as *cliques* in $\mathcal{G}$ (also called *groups* in what follows). One advantage of this approach is its parsimony. Since clusters define sets of nodes with strong interaction patterns, a clique seems here to be the most suitable entity because it requires that all its elements have to be linked to all the others. Beside being the most constrained definition, cliques also define the smallest cohesive units in a graph. Since the phylomemy workflow builds higher order clusters from the initial clusters, taking into account both the temporal dynamics and the $\mathcal{G}$ structures, the choice of cliques as basic clusters will allow us to obtain the finest

possible resolution thereafter. Last, since cliques overlap, the different meanings of a given term can be represented by its participation to different cliques. Consequently, clique detection is the most parsimonious choice that offers the largest latitude to achieve multi-level temporal reconstruction. Clique detection generates a very high number of clusters, some being included in others which produces a form of redundancy. Moreover, given that $\mathcal{G}$ is a weighted graph, it is possible to further characterize cliques by means of indicators based on the strength of their links that reflect the attention (in terms of number of papers produced) of the scientific community for the corresponding topics. It has already been shown that such indicators could be correlated with the fate of a research domain in a phylomemy . The set of all possible cliques can consequently be filtered to improve the degree of informativeness and accuracy of a phylomemy. This filtering also helps to make the full workflow more tractable. The output of ${}^3\Phi$ is thus a set of roots clusters computed over $\mathcal{T}^*$: $C^* = \{C^T | T \in \mathcal{T}^*\}$ with $C^T = \{C_j | j \in J^T, \}$ where $C_j = \{r_i \, | i \in \mathcal{I}_j \ \subset \mathcal{I}\}$. We will note $C = \bigcup_{C^T \in C^*} C^T$ the set of all clusters over all periods.

Implementation of ${}^3\Phi$ in Gargantext Within each period, in order to turn $\mathcal{G}$ and $\mathcal{W}$ into sets of groups $C^T = \{C_j | j \in J^T, \}$, we let the user choose between two clustering techniques:

1. The Frequent Item Sets (FIS). Given a minimum frequency threshold, FIS are here sets of n-grams that jointly occur in a minimum number of documents (called *support*). To that end, we make use of the Haskell's implementation[29] of the *Linear timeClosed itemsetMiner* algorithm (LCM) . The resulting FIS are then filtered by a minimum value of support and *size* (i.e., number of n-grams in a FIS). We finally remove the FIS that are included within larger ones. If a given period eventually ends up without any FIS, we lower both the support and the size until we succeed in repopulating it. We thus aim to avoid reconstructing hollow phylomemies.
2. The Maximal cliques (MC). In a given graph, MC are complete sub-graphs of co-occurring n-grams that cannot be extended by adding one or more adjacent nodes. In our case, we start by filtering the conditional probability graph either with a fixed threshold on the probability values or by selecting a maximal number of out-neighbors per n-grams. We then make use of Gargantext's MC detection technique . The resulting cliques are finally filtered by size and by following an equivalent repopulation mechanism for empty periods.

As Fis have the particularity to be directly linked to a number of documents of appearance (by means of their support), using them to build a phylomemy will produce more generic and lately established scientific fields. On the contrary, the MC will capture early dynamics through broad associations of similar papers. Once the fields are shaped, we include them, period by period, in our phylomemetic data structure. The phylomemy is now made of orphans timestamped groups of n-grams.

D.4 Inter-temporal matching (${}^4\Phi$)

Once the fields of knowledge have been identified for a given spatio-temporal resolution, the operator of inter-temporal matching ${}^4\Phi$ reconstructs the lineage between these fields. It assigns to every field a set of parents and children that highlights a continuity over time of the research questions or objects of study addressed by the corpus. The output structure is a phylomemy that can be further explored, quantified and visualized. An upstream only inter-temporal matching might miss some splitting events where one field gives birth to several others. For this reason, in the whole workflow we will do both upstream and downstream inter-temporal matching to optimize the identification of both merging and splitting events in a phylomemy. We consequently consider the function ${}^4\Phi^{\succ}_{\delta} : \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(C), w)$:

$$ {}^4\Phi^{\succ}_{\delta}(C^T) = (\{C_j \in \kappa^{\succ}_{C^T}\}, w) $$

Where:

- $\kappa^{\succ}_{C^T} = argmax_{\Delta(C^T,\mathcal{K})}[\arg min_{\{\kappa \subset C^{T' \succ\succ T} | \Delta\left(C^T,\mathcal{K}\right) \geq \delta\}} \{\tau(C^T, \kappa)\}]$,
- $C^{T' \succ\succ T} = \{C^{T'} \in \mathcal{C} | T' \succ\succ T\})$ is the set of all clusters of $\mathcal{C}$ which period is strictly a posterior to T,
- $w \in [0,1]$ is the association strength of $\kappa^{\succ}_{C^T}$,

By combining upstream and downstream inter-temporal matching, we obtain the symmetric function ${}^4\Phi : \mathcal{C} \times [0,1] \longmapsto (\mathcal{P}(C), w)^2$ defined as: ${}^4\Phi_{\delta}(C) = (({}^4\Phi^{\prec}_{\delta}(C), {}^4\Phi^{\succ}_{\delta}(C))$.

[29] See https://lig-membres.imag.fr/termier/HLCM/hlcm.html

Implementation of ${}^4\Phi$ *in Gargantext* The operator of inter-temporal matching ${}^4\Phi$ aims to weave kinship connections between the groups from one period of time to another and for a given similarity threshold $\delta \in [0, 1]$. We process it from the groups of the most recent periods to the groups of the oldest ones and, for each of them, we successively try to find both upstream (i.e., the parents) and downstream (i.e., the children) inter-temporal links able to satisfy δ. Within a given period T_i, we start by framing the ranges of upstream/downstream matching periods $[T_{i+1}, T_{i+t}]$ and $[T_{i-1}, T_{i-n}]$ by taking into account a *time window* parameter $t, t \geq 1$. We gather their corresponding groups and reduce them as a list of possible parents/children candidates. Thus, any parent candidate $C_k, k \in [T_{i+1}, T_{i+t}]$ will have to share at least one n-gram with the current targeted group $C_j, j \in T_i$. The matching strategy now consists in finding kinship connections between C_j and singles or pairs of candidates from T_{i+1}. If no singles/pairs from T_{i+1} are strong enough to satisfy δ, we extend the making of the singles/pairs to the candidates from $[T_{i+1}, T_{i+2}]$ and so on until $[T_{i+1}, T_{i+t}]$. In a pair of candidates, at least one the two groups must come from the oldest period of the current time window. We here make use of the inter-temporal matching function ${}^4\Phi_\delta(C)$. Once one or more singles/pairs of parents/children have satisfied the threshold δ, we sort and group them by decreasing value of similarity in order to only keep the strongest set of parents/children. We repeat this operation for all the targeted groups and add the resulting kinship connections in our data structure as sets of weighted pointers. The *branches* are finally identified and named by looking for all the newly shaped connected components of groups. A reference to their belonging branch is set up within the groups.

D.5 The sea elevation and its implementation in Gargantext

The operator ${}^4\Phi$ is managed by a *sea level* mechanism detailed in 3.2.4. There, regarding an initial level of similarity δ_0 and a fixed increment, ${}^4\Phi$ is recursively and locally processed within each of the branches sorted by size. An initial inter-temporal matching task is first fulfilled at δ_0 to set up the sea level algorithm. The latter is next executed for increasing value of δ while the outcomes of ${}^4\Phi$ satisfy the overall objective function $F_\lambda(\varphi)$ described in 3.2.3 and parameterized by the level of observation $\lambda \in [0, 1]$. After each elevation, we update the branches names by re-indexing the steady ones and framing the emerging ones. Branches references within the groups are updated as well.

D.6 Alternative choice function Ξ

As described in section 5.2, the choice functions Ψ and Ξ that model the users' preferences are key building blocks of the quality function F_λ. In order to explore the effects of changing these functions, we have experimented an alternative Ξ that weights a term x by its frequency p_x in $\mathcal{D}$.

With this new Ξ, the objective function $F_\lambda(\varphi)$ described by Equation 2 can be written as follows (Equation 6):

$$F_\lambda(\varphi) = \Sigma_{x \in \mathcal{L}} \frac{p_x}{\Sigma_{y \in \mathcal{L}} p_y} . \Sigma_{B_k \in \mathcal{B}^x_k} \frac{|B_k|}{\Sigma_{B_j \in \mathcal{B}^x_k} |B_j|} . F_\lambda(x, k) \tag{6}$$

Applying the same analysis as in 4.1.2 with the parameters of Table 4 we obtained the results described in Figure 24. These new results confirm that the sea-level rise algorithm outperforms the original method.

A qualitative analysis of the phylomemies obtained by this variant led us to conclude that this alternative Ξ function leads to qualitatively less informative phylomemies than the ones reconstructed with Equation 4 and reviewed in the core of this article. Our hypothesis is that, from the cognitive points of view, interesting terms are not the most frequent ones nor the least frequent ones, but those that have intermediate level of frequencies. We conjecture that the most meaningful phylomemies for a given community of users will be reached by choosing Ξ as suggested in section 5.2, *i.e.* through the analysis of the queries frequencies in a scholar database ; and that doing so, Ξ should display a bell shape with respect to terms frequencies.

D.7 The ghost ancestors and their implementation in Gargantext

The ghost ancestors mechanism is introduced in 3.2.6 for reuniting branches heads and reducing artificial branches splits. For a given branch B^k, the latter mechanism relies on the last level of elevation δ_{B_k} reached by the branch's head C^T (i.e., the oldest group of B^k). In order to do this, we manage to keep on the fly a trace of the branches drifting history within the data structure of the groups throughout the whole sea elevation process and evolution of δ. After having found ancestors, we store the resulting ghost pointers in the phylomemy and re-index the new branches names and references if needed.

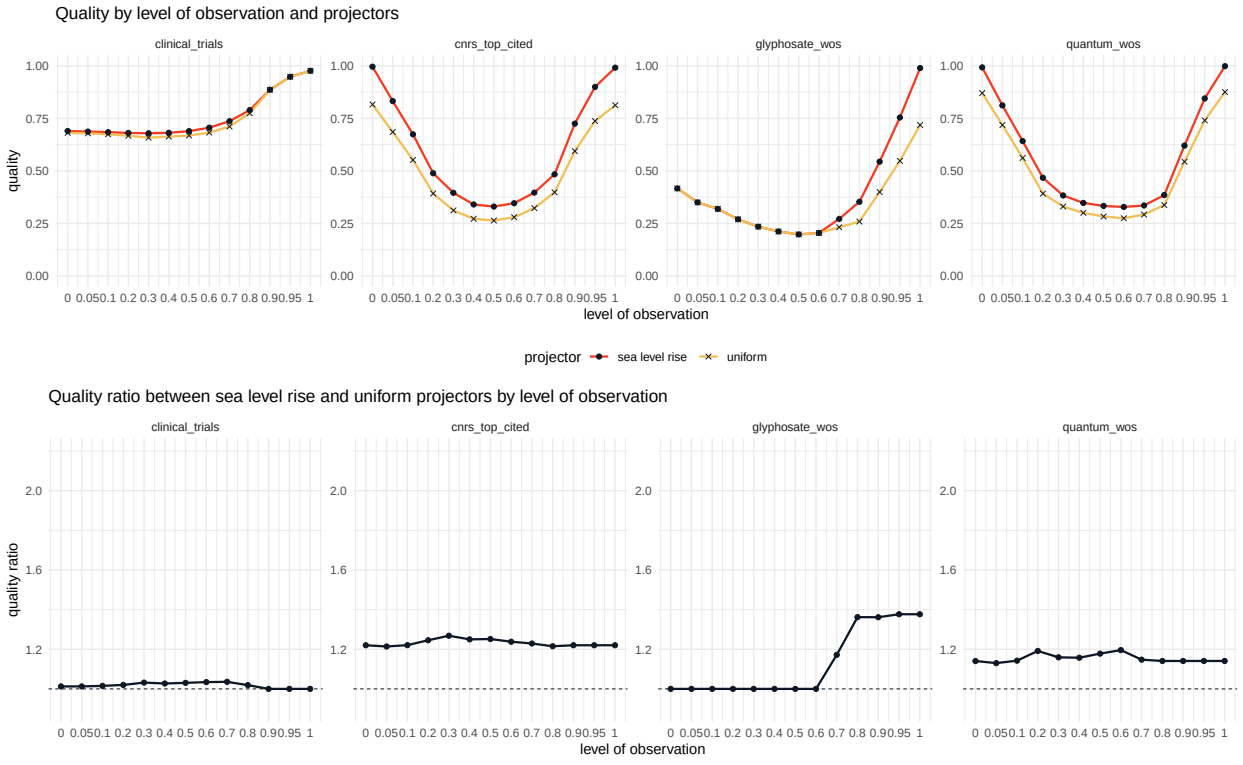

Figure 24: Comparison between the sea-level rise and the uniform projectors for four distinct corpora with the objective function of Equation 6.

D.8 The scales of observation and their implementation in Gargantext

In order to take into consideration the *scales of observation* investigated in 3.2.5, our implementation let the user choose between two approaches. The first one relies on the sea level mechanism and consists in re-processing it until $\delta = 1$ on the basis of a phylomemetic network φ obtained in D.5. Within each of the branches, groups connections will thus break up once again and give us a new drifting history related to specific values of δ. The latter are then stored within the data structure of the groups and eventually used for zooming purposes in a dedicated visualization system . The second one is based upon a *synchronic clustering* process. By using the matching function of **??**, we here weave new kinship connections between groups that belong to a shared period[30]. We end up with a synchronic graph of similarity per period that we later prune by means of a similarity threshold parameterized by the user. We finally compute the remaining connected components and turn them into new larger groups. This second approach leads to branch mergers.

D.9 Post-processing and its implementation in Gargantext

The various post-processing tasks aim to prepare a given phylomemetic network φ for its exportation and eventually its visualization. For each of the branches obtained in D.5, we first place ourselves at the height of their last local δ elevation and select the corresponding groups and kinship connections. The latter will next go through four subsequent operators:

1. The metrics. Here we enrich the groups with various dynamical information. First, we determine whether a given n-gram is *emerging* (i.e., used for the first time in φ) or *decreasing* (i.e., used for the last time in φ). Next, we compute for each ngrams their scores of *genericity*, *specificity* and *inclusion* as well as a measure of *tf-idf* calculated within their respective branch. We also enrich the whole branches with information about their age, date of birth and size.

[30]The user can here choose to process the synchronic clustering within the scope of single branches or between sibling branches (i.e., branches that split at a shared similarity level δ_{B_k}) or between all the branches.

2. The labeling. Here we label the branches and the groups to give them a meaningful and human readable name. For a given group, the label is made of its two most inclusive n-grams formerly selected among the emerging ones if they exist. For a given branch, the label is made of the two n-grams that have the highest *tf-idf* score and also formerly selected among the emerging ones.

3. The sorting. Here we sort the branches by making use of the hierarchical clustering inherited from 3.2.4. By doing so, we bring similar branches closer and move the ones that split at a low level of δ away. Closely related siblings branches are finally sorted by ascending date of birth.

4. The filtering. Here we filter the smallest branches in order to clarify the upcoming visualization of φ. We thus let the user select a minimum number of periods and we then move away the branches that do not cover a time window superior or equal to the this threshold.

D.10 Export and its implementation in Gargantext

The phylomemetic network φ is now ready to be exported and recorded on a file. We here choose to rely on the spatialization features of *Graphviz*[31] to render our phylomemy. We thus start by defining a main graph object enriched with general information like the original number of documents, the number of exported groups, etc. We then create a sub-graph for each of the remaining branches that we later populate with their respective groups, n-grams and kinship connections turned into Graphviz objects. Dynamical metrics computed in D.9 are here embedded as ad hoc attributes. We finally export our phylomemy in the form of a Graphviz Json file suitable for future visualization tasks .

[31] See `https://www.graphviz.org/`