Scientometrics (2022) 127:545–575
https://doi.org/10.1007/s11192-021-04186-5

Draw me Science

Multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies

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

Received: 10 April 2021 / Accepted: 30 September 2021 / Published online: 22 November 2021

Abstract
In 1751, Jean le Rond d'Alembert had a dream: "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". In this paper, we address the question identifying the branches of science by taking advantage of the massive digitization of scientific production. In the framework of complex systems studies, we first formalize the notion of level and scale of knowledge dynamics. Then, we demonstrate how we can reconstruct a reasonably precise and concise multi-scale and multi-level approximation of the dynamical structures of Science: phylomemies. We introduce the notion of phylomemetic networks—projections of phylomemies in low dimensional spaces that can be grasped by the human mind—and propose a new algorithm to reconstruct both phylomemies and the associated phylomemetic networks. This algorithm offers, passing, a new temporal clustering on evolving semantic networks. Last, we show how phylomemy reconstruction can take into account users' preferences within the framework of embodied cognition, thus defining a third way between the quest for objective "ground truth" and the ad-hoc adaptation to a particular user's preferences. The robustness of this approach is illustrated by several case studies.

Keywords Phylomemy reconstruction · Knowledge dynamics · Phenomenological reconstruction · Multi-scale and multi-level complex systems · Science map · Co-word analysis

Mathematics Subject Classification 05C82 · 91C20 · 91D30

JEL Classification D85

✉ David Chavalarias
david.chavalarias@iscpif.fr

[1] CNRS, Complex Systems Institute of Paris Île-de-France (ISC-PIF), 113 rue Nationale, 75013 Paris, France

[2] EHESS, Centre d'Analyse et de Mathématique Sociales (CAMS), 75006 Paris, France

Introduction

The shapes of science

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, many disciplines have sought to produce a comprehensive representation and understanding of the knowledge mankind has produced to date, from the history and philosophy of science (Chavalarias et al. 2021) to the emerging field of science mapping (Chen 2017).

Science can be defined quite generically as "(1) body of knowledge, (2) method, and (3) way of knowing" (Abell and Lederman 2007). This body of knowledge, resulting from the distributed interactions of thousands of scientists over the years, is nowadays almost entirely digitized, making large-scale quantitative analysis possible.

In this article, we will give a formal definition to the notion of "branch of knowledge" in the framework of co-word analysis. We demonstrate how we can reconstruct, from its massive body of knowledge, a reasonably precise and concise approximation of the structure of Science that can be grasped by the human mind and explored interactively (an operation called *phenomenological reconstruction*[2] (Bourgine et al. 2009; Chavalarias et al. 2021). We will then apply this approach to several case studies.

Levels and scales in knowledge dynamics

Scientific research domains are sustained by entangled socio-economic processes that guide the progress of science. Such complex systems display structures at all scales embedded in a hierarchical organization (Chavalarias 2020). 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 (Li et al. 2005).[3]

The method presented in this paper makes a clear distinction between these notions of *level* and *scale* in the phenomenological reconstruction of knowledge dynamics.[4] 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.[5] 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

[1] "Former un Arbre généalogique ou encyclopédique qui les rassemble sous un même point de vûe, & qui serveà 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.

[3] In biology for example, the choice of level of observation determines what the main entities under study (organs, cells, genes, etc.) are, while the choice of a scale determines the smallest resolution adopted to describe these entities.

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

[5] 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.

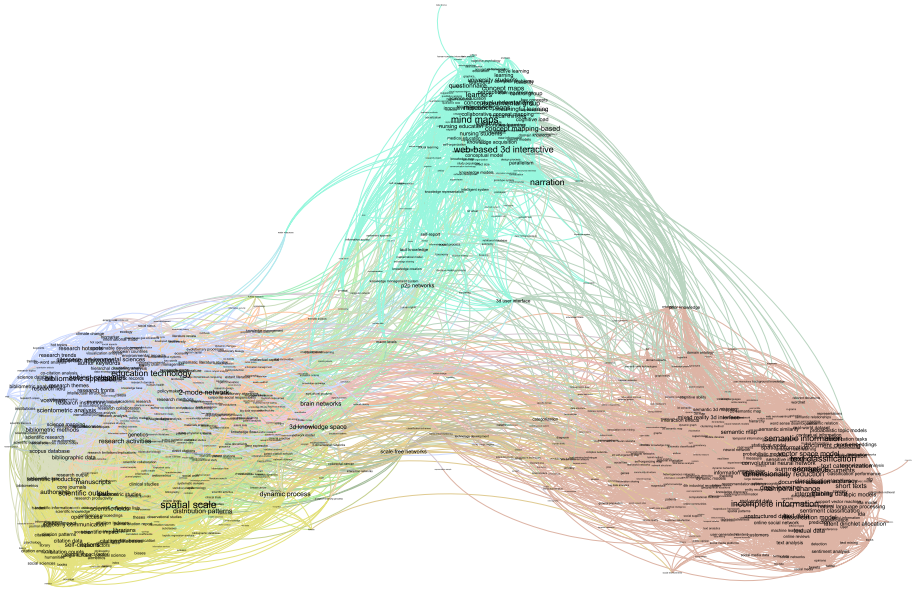

Fig. 1 Map of the domain of *science and knowledge mapping* ($\mathcal{D}_{maps}$: 13,844 related publication meta-data extracted from the Web of Science (see SI B for details). This map reveals the three main research areas concerned with knowledge mapping: bibliometrics (on the left), information retrieval & documents classification (on the right) and education science (at the top). Generated with Gargantext and spatialized in Gephi (Bastian et al. 2009) with the Force Atlas algorithm (Jacomy et al. 2014). An interactive version of this map is available at http://maps.gargantext.org/maps/sciencemaps and a stand alone version can be downloaded from David Chavalarias and Delanoe (2021)

derive their unity from some underlying dynamic proces–; while the notion of scale does not necessarily imply time. We propose thereafter a quantitative definition of these notions of *level* and *scale* of observation.

Related work

Mapping science and knowledge

In order to identify the main areas concerned by the domains of *science mapping* and *knowledge mapping*, we have performed a co-word analysis on a corpora $\mathcal{D}_{maps}$ of 14k related documents (see SI B for details). Figure 1 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 for qualifying the underlying socio-semantic structures,
- The *information retrieval* and *documents classification* approaches (on the right) aim at finding optimal methods for interacting with large corpora of publications,
- The *education science* and *cognition* approaches (at the top) focus on methods that enhance the capacities of learners.

The question of mapping the dynamics of knowledge has been mainly addressed by the first two research areas that include sub-areas described by Fig. SI 2, each of them having known recent developments regarding evolution, dynamics and temporality concerns.

1. **Bibliometrics** is divided into two major components:
   - *Citation and co-citation analysis (CA)* Primarily focused on the assessment of scientific output (Garfield 1972), it quickly diversified with the analysis of large citation landscapes through methods such as *co-citation* and *bibliographic coupling* (Kessler 1963; Small 1973). Later on, following the creation of the *Web*, these methods were generalized as part of the study of *hyperlinked data* (Kleinberg 1999). Methods to describe *conceptual structures* of science such as *research fronts*, *hot topics* and *trends*, etc. White and McCain (1998) and Börner et al. (2003) 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 (Chen 2006; Claveau and Gingras 2016; Cambe et al. 2020).
   - *Co-word analysis (CWA)* is a bottom-up approach first developed by sociologists in the 1980s (Callon et al. 1986) 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 (Chavalarias and Cointet 2013; Wang et al. 2014; Rule et al. 2015). This methodology has the advantage of being suitable to any kind of textual corpora regardless of structure.

   Citation and co-word analysis research areas share common objectives, namely to understand the structures of a body of knowledge. They form a toolkit that is at the core of the *science of science*. Their parallel development has quickly paved the way to hybrid research between co-word and citation analysis (Braam et al. 1991; Boyack and Klavans 2010) or social and semantic networks (Roth and Cointet 2010). Finally, the growth of scientific databases has stimulated the visualization of wide *citation landscapes* (Small 1997) or complex atlases of sciences (Börner 2010, 2015).
2. **Information retrieval (IR)** has mobilized latent semantic analysis and topic modeling in the early 2000s at the instigation of a community of statisticians (Blei et al. 2003). One of the aim was to adopt a "bottom-up" approach to structure *collections of documents* retrieved by queries. The use of topic modeling mostly focuses on *document classification* (Wei and Croft 2006), *recommendation* (Wang and Blei 2011) or *sentiment analysis* (Lin and He 2009). Some scholars also proposed methods to organize a set of retrieved documents according to their topics and temporal relationships (Liao and Qian 2019; Shahaf et al. 2012). More recent works started to tackle the mapping of scientific issues and their dynamics at the scale of a research domain (Millar et al. 2009; Gohr et al. 2009; Cui et al. 2011; Shahaf et al. 2013; Yang et al. 2017; Jähnichen et al. 2018).

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* (Tshitoyan et al. 2019; Tacchella et al. 2020). It has been mostly used for the purpose of information retrieval and documents classification (which is the reason why it appears in the same group as LDA in Fig. 1), but can also be a useful tool to analyze science's evolution, mostly at the micro-level of terms-to-terms similarities (Palmucci et al. 2019; Tshitoyan et al. 2019; Tacchella et al. 2020).

For this paper, we have chosen not to make any assumptions about the structure of the data and to place ourselves in the framework of co-word analysis (Callon et al. 1983) where the only requirement is the existence of timestamped text strings. When other meta-data are available, this method will definitely benefit from being complemented or hybridized with other approaches such as citation analysis and co-author analysis.

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. Finally, 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* (Bourgine et al. 2009; Chavalarias et al. 2021). 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]

In this paper, we will work with a methodology called *phylomemy reconstruction* (Chavalarias and Cointet 2013).

We will define formally a category of meanings conveyed by *phylomemies* (Chavalarias and Cointet 2013) as phenomenological reconstructions. We will then show how this definition can 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) $\mathcal{R} \longmapsto \mathcal{V}$, the projector that selects a level of observation for multi-scale exploration. At the same time, we make tangible the different shapes generated by science, which can be visualized with an appropriate free software (cf. Lobbé et al. 2021). 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 Chavalarias et al. (2021).

Materials and methods

Generic workflow of phylomemy reconstruction

The workflow of phylomemy reconstruction (cf. Fig. 2 and Chavalarias and Cointet 2013) 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 (Fig. 3) and is detailed as a case study in "Reconstruction of the history of a research domain: the example of glyphosate research" section.

[6] Even though assumptions 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 assumptions.

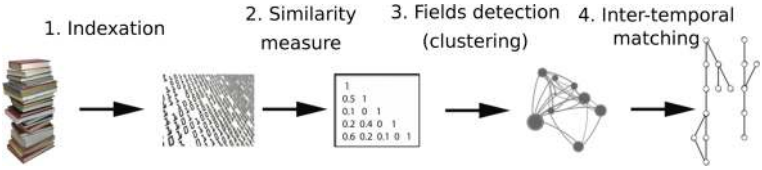

Fig. 2 Workflow of phylomemy reconstruction from raw data (digitized textual corpora) to global patterns. The output is a set of phylomemetic branches where each node is constituted by a network of terms describing a research field. These nodes are a proxy of scientific fields and can have different statuses: emergent, branching, merging, declining. *Source*: Chavalarias and Cointet (2013)

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

1. **Indexation.** The core vocabulary of $\mathcal{D}$ is extracted as a list $\mathcal{L} = \{r_i \,|i \in \mathcal{I}\}$, 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. **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 by the research question at stake, as described for example in Dias et al. (2008) or Weeds and Weir (2005). Depending on the question, two reconstructions using two distinct classes of similarity measures may prove complementary (see also SI. B).
3. **Fields detection and clustering.** Within each period, the completion of ${}^2\boldsymbol{\Phi} \circ {}^1\boldsymbol{\Phi}$ results in time 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 multi-terms characterizing key research questions. Research on community detection algorithms has been very prolific (Fortunato 2010) and several algorithms with different intrinsic spatial resolution could potentially be suitable at this step. The result of fields detection is a *time series of clustering* $\mathcal{C}^*$ on roots C_j computed over $\mathcal{T}^*$: $\mathcal{C}^* = \{\mathcal{C}^T | T \in \mathcal{T}^*\}$ where $\mathcal{C}^T = \{C_j | j \in J^T, \}$ and $C_j = \{r_i \,| r_i \in \mathcal{L}, i \in \mathcal{I}_j \subset \mathcal{I}\}$. We will note $\mathcal{C} = \bigcup_{\mathcal{C}^T \in \mathcal{C}^*} \mathcal{C}^T$ the set of all clusters over all periods.
4. **Inter-temporal matching.** Once the fields of knowledge have been identified as a temporal series of clusters, inter-temporal matching reconstructs the lineage between these clusters. This inter-temporal matching operation brings time into play. Consequently, this is where the notion of *level* arises. The result of ${}^4\boldsymbol{\Phi}$ is an evolving structure describing the evolution of large knowledge domains and it has been called a *phylomemetic network* (Chavalarias and Cointet 2013).

[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. *{climate change, global warming}*.
[8] Also called *graph clustering* algorithm.

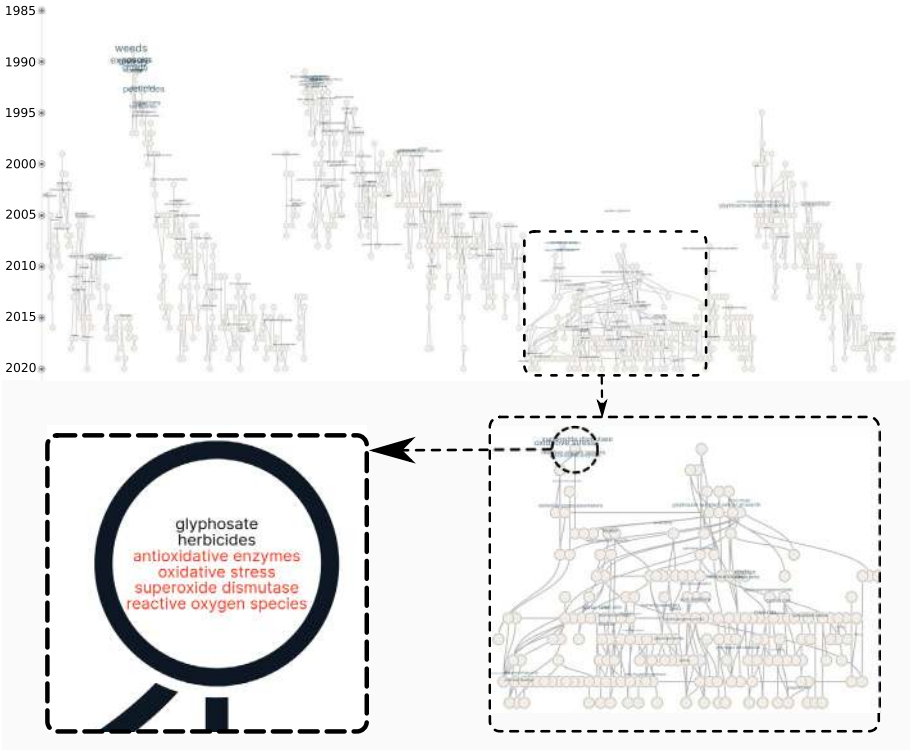

Fig. 3 The visualization of the phylomemy $\mathcal{D}_{glyphosate}$ (16,655 documents) at the level $\lambda = 0.8$ between 1995 and 2020 with branches smaller than 3 filtered out. Each connected dot represents a specific field of research described by several key-words (as displayed for example in the magnified area. 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. Details about this case study are given in "Reconstruction of the history of a research domain: the example of glyphosate research" section. An interactive version of this visualization, available online at http://maps.gargantext.org/phylo/glyphosate, can be downloaded from the archives (David Chavalarias and Delanoe 2021) (data), (Quentin Lobbe and Chavalarias 2021) (explorer). See Lobbé et al. (2021) for details

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 ${}^{4}\boldsymbol{\Phi}$ and how it can convey the notions of *level* and *scale*, setting arbitrary parameters for ${}^{1}\boldsymbol{\Phi}$ to ${}^{3}\boldsymbol{\Phi}$. The reader interested in these particular steps can find more examples and technical details in Chavalarias and Cointet (2013), Cointet and Chavalarias (2008) and Chavalarias and Cointet (2008).

[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).

Method of reconstruction of the dynamics

The most general assumption on the evolution of fields of knowledge, subsequently defined has roots clusters, is that some are likely to emerge, split, merge or die (Palla et al. 2007; Rossetti and Cazabet 2018). 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 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 (the children of C^T). This is achieved through the definition of some parentage metrics and the selection of the most relevant 'parents' and 'children' for each cluster, when they exist. From a conceptual point of view, we would like these inter-temporal matching to allow us to define lineages of scientific fields that would coherently describe the evolution of scientific knowledge. These lineages would constitute the *branches of science*.

There is however a subtlety here. Once the parentage metrics is chosen, computing the weighted inter-temporal associations between clusters is the easy part. The hard part is to choose which ones to keep.

First, since filtering out some inter-temporal associations has a direct impact on the overall connectivity of the dynamic structure, and consequently on the number of branches, the interpretation of the final result depends strongly on the pruning procedure.

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. The trade-off between pruning weak inter-temporal associations and highlighting continuities in evolution is not straightforward. When allowing matching between non-consecutive periods, this trade-off directly influences the average time difference between related fields and the global understanding of the final output.

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. We will now introduce the main concepts that will be used to formalize the notions of *levels* and *branch of science*.

Upstream and downstream inter-temporal matching

Let $\Delta : \mathcal{C} \times \mathcal{P}(\mathcal{C}) \rightarrow [0, 1]$ be a similarity measure 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] Chavalarias and Cointet (Chavalarias and Cointet 2013) 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' ${}^4\boldsymbol{\Phi}^{\prec}_{\delta}(C^T)$ according to Δ:

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

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,

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

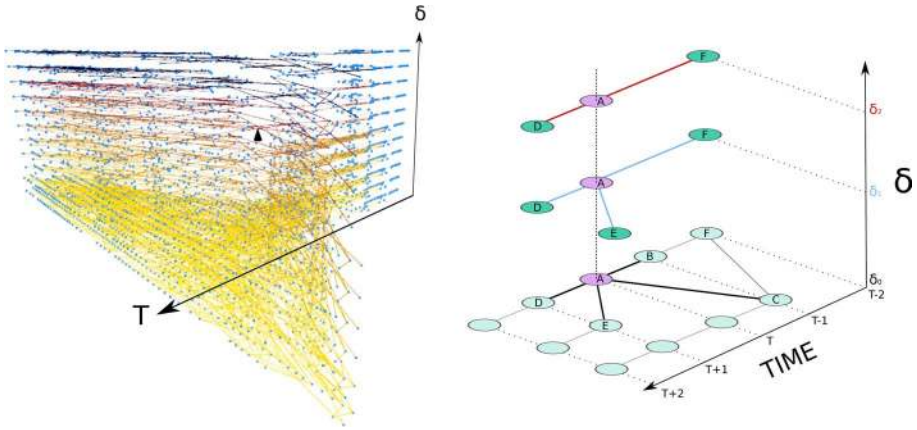

Fig. 4 Foliation on a temporal series of clustering and its local structure. *On the left* a foliation on a temporal series of clustering (based on $\mathcal{D}_{maps}$ see also Fig. 9) parameterized by δ (δ values have been discretised for the plot). Each dot is a cluster. *On the right* Local structure. Above the cluster $A \in \mathcal{C}^T \subset \mathcal{C}^*$ 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 {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 $\mathcal{C}^*$

- $\tau(C^T, \kappa)$ 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'.

As the definitions of parents and children of a cluster are symmetrical with respect to time, we will also consider in this paper the symmetric down-stream inter-temporal matching function ${}^4\boldsymbol{\Phi}^{\succ}_{\delta}(C^T)$, 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 C.4 for full description).

We thus define ${}^4\boldsymbol{\Phi} : \mathcal{C} \times [0, 1] \longmapsto (\mathcal{P}(\mathcal{C}), w)^2$ as:[12] ${}^4\boldsymbol{\Phi}_{\delta}(C) = (({}^4\boldsymbol{\Phi}^{\prec}_{\delta}(C), {}^4\boldsymbol{\Phi}^{\succ}_{\delta}(C))$

Phylomemies as foliation on time series of clustering

Thereafter, we will work with the following definitions:

[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.

[12] 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 "Comparison with previous works" section that are going beyond the simple correspondence between consecutive periods.

Definition *Time 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 time series of clustering over $\mathcal{L}$ is defined as $\mathcal{C}^* = \{\mathcal{C}^T | T \in \mathcal{T}^*\}$ where $\mathcal{C}^T = \{C_j^T | C_j^T \subset \mathcal{P}(\mathcal{L})\}_{j \in J^T}$ is a set of clusters of elements of $\mathcal{L}$

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

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

Definition *Phylomemy*. A *phylomemy* ϕ is a foliation on a temporal series of clustering $\mathcal{C}^*$ (cf. Fig. 4). It describes, for any cluster C_j^T in temporal components of $\mathcal{C}^*$ and *any* threshold δ, the relevant inheritance linkages of C_j^T. Thereafter, we will consider the space of all foliations on temporal series of *roots* clustering as the space $\mathcal{R}$ for the study of knowledge dynamics.

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 $\mathcal{C} = \bigcup_{\mathcal{C}^T \in \mathcal{C}^*} \mathcal{C}^T$ such as $\forall C_j^T \in \mathcal{C}; \varphi(C_j^T)(1,1) \subset \mathcal{P}(\mathcal{C}^{T' <<^T})$ and $\varphi(C_j^T)(2,1) \subset \mathcal{P}(\mathcal{C}^{T' >>^T})$.[13]

Definition *Phylomemetic network*. Let ϕ be a phylomemy over $\mathcal{C}^*$ and $\Pi : \mathcal{C} \longmapsto [0,1]$, a *phylomemetic network* is defined by $\varphi_\Pi = \{(C_j^T, {}^4\boldsymbol{\Phi}^{<}_{\Pi(C_j^T)}(C_j^T)) | C_j^T \in \mathcal{C}\}$. It is a *plaque* of the phylomemy ϕ that defines a weighted inheritance networks over a temporal series of root clusters.

Definition *Branches of a phylomemetic network*. Let φ be a *phylomemetic network*. It can be written $\varphi = \bigcup_k B_k$ where each B_k is a connected component of the temporal network formed by the inter-temporal links. $\{B_k\}_k$ will be called the *branches* of φ. It is thereafter possible to visualize theses branches in $\mathcal{V}$ in order to understand the structure of ϕ (cf. Fig. 3 and Lobbé et al. 2021).

In this paper, $\mathcal{V}$ will be 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. But as there is a huge number of plaques, a central question comes out for further analysis: How to find the most meaningful ones for *us*.

These definitions generalize the work of Chavalarias and Cointet (2013) who appear to have studied the specific case where the operators $\mathcal{R} \longmapsto \mathcal{V}$ are all uniform projectors $\Pi = \delta$, *i.e.* $\varphi_\delta = \{(C_j^T, {}^4\boldsymbol{\Phi}^{<}_{\delta}(C_j^T)) | C_j^T \in \mathcal{C}\}$.

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

Branches of knowledge and levels of observation

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 the question $\mathcal{Q}(x)$: "Can you show me a branch of knowledge that deals with x?"

From the perspective of a phylomemy of science $\phi \in \mathcal{R}$, the question can be rephrased as follows: (1) which phylomemetic network φ of ϕ should we choose? (2) which branch of φ should we propose to d'Alembert?

We cannot answer these questions with the approach of Chavalarias and Cointet (2013) because there is no indication on how to choose the appropriate uniform projector $\Pi = \delta$. The same can be said from other contributions dealing with re-emerging topics, such as Jo et al. (2011) who are at a loss when it comes to setting an inter-temporal matching threshold.

Moreover, there is no reason to think that a uniform projector Π_δ will provide the same level of observation for all branches of knowledge. More likely, the branches of knowledge at a given level of observation could differ with respect to their minimal inter-temporal matching threshold.

To overcome these issues, we can draw inspiration from information retrieval. We can propose two metrics to assess the relevance of a branch B_k to $\mathcal{Q}(x)$:

- 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. It is related to the probability to observe x by choosing at random a cluster in B_k within $\mathcal{T}_x$,
- The *recall* $\rho_x^k = \frac{|\mathcal{C}_x \cap \mathcal{C}_{B_k}|}{|\mathcal{C}_x|}$ of B_k against x. It is related to the probability to be in B_k when choosing a cluster about x at random in ϕ.

where:

- $\mathcal{C}_x$ is the set of all fields of φ containing x,
- $\mathcal{T}_x$ are the periods covered by $\mathcal{C}_x$,
- $\mathcal{C}_{B_k}$ is the set of all the fields of the branch B_k

An answer B_k to $\mathcal{Q}(x)$ 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, it appears that D'Alembert must also to indicate the desired 'trade-off' in order for us to answer his question.

Let's define this trade-off by a variable $\lambda \in [0\ 1]$, we can evaluate the quality of an answer $\mathcal{Q}(x)$ 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 $\lambda = 1$, only the recall counts.[14]

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

Several branches could mention x, which means that we also have to know which B_k 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 proposed first to D'Alembert.

We thus obtain an objective function that evaluates the relevance of φ in answering $\mathcal{Q}(x)$:

$$F_\lambda^x(\varphi) = \Sigma_{B_k \in \varphi | B_k \cap \mathcal{C}_x \neq \varnothing} \Psi_x(k).F_\lambda(x,k) \tag{1}$$

We could search, for each question $\mathcal{Q}(x)$, 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, we thus have to assess the relevance of a particular phylomemetic network for its ability to answer *any* question $\mathcal{Q}(x)$ 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.

The global F-score of a representation φ of a phylomemy is then defined as:

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

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 "Discussion" section how these 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 is set 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 phylomemetic networks with the highest F_λ score will be the ones that include large complex branches whereas for low λ values, those 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.

Adaptive inter-temporal matching and step phylomemetic networks

For any desired scale of observation λ of $\phi \in \mathcal{R}$, we can now evaluate any projection φ in $\mathcal{V}$ with $F_\lambda(\varphi)$.

Previous works on phylomemy reconstruction have so far only taken into account uniform projectors. Here, we propose to consider a new class of adaptive projectors defined over $\mathcal{R}$ and parameterized by the level of observation λ. They are designed to map the internal dynamics of each branch of science and thus outperform uniform projectors. To that end, we will need the following definitions:

Definition *Uniform step projector*. 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 $\mathcal{C}^*$ 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 *Step phylomemetic network*. 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 $\varphi = {}^4\Phi_\Pi(\phi)$. The family of step phylomemetic networks extends the family of phylomemetic networks obtained by mean of uniform projectors.

Step phylomemetic networks are phylomemies' projections of particular interest regarding the homogeneous conception of what inheritance means within each of their branches (all inter-temporal links have been processed with threshold $\delta^\lambda_{B_k}$ while this threshold can differ from one branch to another). This property makes it possible to take into account the internal dynamics of each branch and makes it easier to interpret their morphologies. What remains now is to design an algorithm capable of finding an optimal *step phylomemetic networks* for a given level of observation λ.

Sea-level rise algorithm

In order to find a step phylomemetic networks optimized for a given level of observation, let's start by noticing 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,[15] 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 $\Pi = 0$ associated to a phylomemetic network noted $\varphi_0 \in \mathcal{V}$.

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 $\mathcal{C}^{B_k} \subset \mathcal{C}^*$ of φ_0.

After each local level rise, F_λ is used to evaluate the validity of this level's increase and, in case of a branch split, the subsequent level increases are handled independently within each of the resulting branches (Fig. 5). Details of this algorithm are given in SI C.5. together with an open source implementation.

By design, the phylomemetic networks generated by the sea level algorithm are *step phylomemetic networks*. Since inter-temporal matching links are recursively reprocessed in the course of this algorithm, it is worth noticing that it may not be possible to transform two observations φ_λ and $\varphi_{\lambda'}$ of ϕ at different levels ($\lambda \neq \lambda'$) simply by pruning the links (see SI G. for an example of a phylomemy described at different levels of observation).

Two levels of observation λ and λ' might therefore convey very different information on the temporal structure of ϕ. In order to fully understand knowledge dynamics, it might be necessary to reconstruct different phylomemetic networks for different values of λ. We present in 4.1.1 a method for determining the preferred level of observation and thus initiate the exploration of a given phylomemy.

[15] 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.

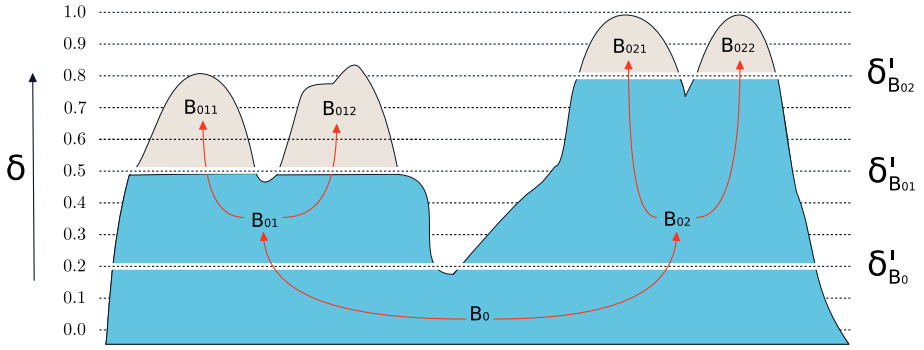

Fig. 5 The elevation of the inter-temporal matching threshold submerges the initial branch $\varphi_0 = B_0$ that first splits into two branches $B_{01} \cup B_{02}$ at δ'_{B_0}; and then each of them splits at different thresholds δ'_{B_01} and δ'_{B_02} to create the final branches of $\varphi = B_{011} \cup B_{012} \cup B_{021} \cup B_{022}$

Phylomemetic networks, hierarchical clustering and endogenous scales of description

Branches of phylomemetic networks φ have complex structures that can be rendered at different scales of description in $\mathcal{V}$ thanks to a synchronic merging of their most related clusters.

The similarity between two clusters is usually evaluated by comparing their components, and clustering methods are numerous. However, since we are dealing with a dynamic structure, we can obtain a hierarchical clustering for free on each $\mathcal{C}^T$ by exploiting the temporal structure of φ. Raising the threshold δ within a given branch without recomputing the parents and children of clusters leads to the progressive pruning of inter-temporal links and to the appearance of new connected components at particular δ values until all this branch's clusters end up isolated. The information about the order of appearance of these connected components can then be used as a basis for a hierarchical synchronic clustering.

This approach makes it possible to both introduce a notion of *scales of description* of a phylomemetic network that fits to the *endogenous structure* of its branches; and define an *endogeneous hierarchical clustering* indexed by these scales of description:

Definition *Scales of description of a phylomemetic branch in* $\mathcal{V}$.

Let B be a phylomemetic branch from 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 inferior to 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 $s^{\delta^{s_B}} \leq card(\mathcal{C}_B)$ 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 $\theta(B, \delta^i)$.

[16] In this operation, contrary to what is made in "Adaptive inter-temporal matching and step phylomemetic networks" section, inter-temporal matching is not reprocessed.

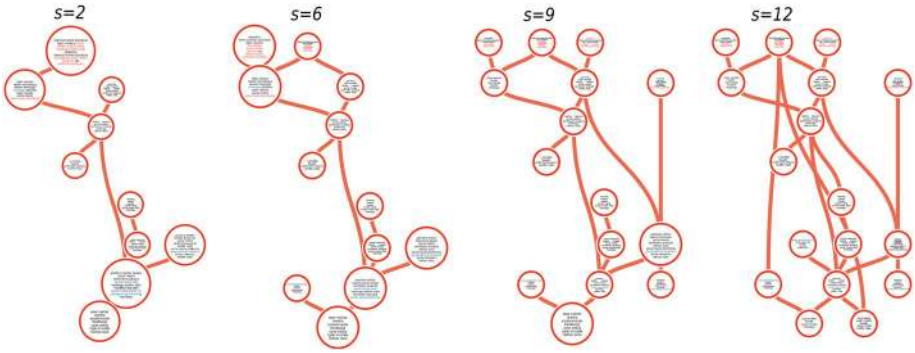

Fig. 6 Endogenous scales of a branch. The branch *Social media/sentiment analysis* of the $\mathcal{D}_{maps}$ phylomemy of Fig. 9 has 12 scales of description. Here we display its internal structure at scales 2, 6, 9 and 12

By construction, $\forall 1 \leq i \leq s_B, \forall C \in \mathcal{C}^T \cap \mathcal{C}_B, \exists! l \in \{1..s^{\delta^i}\} | C \in B_l^{\delta^i}$. 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 $\{\theta(B, \delta^i) \cap \mathcal{C}^T\}_{1 \leq i \leq s_B}$ defines an endogenous synchronic hierarchical clustering over $\mathcal{C}^T \cap B$ indexed endogenously by the scales of description $\{1, \ldots, s_B\}$.

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 $\{1, \ldots, s_{B_k^\lambda}\}$ 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.

An example of a branch described at several scales is given by Fig. 6. 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, each branch has at most one cluster per period corresponding to the synchronic merge of all its clusters of that period. For scale $s > 1$, branches start branching according to their internal structure but remain in constant number. More details on the use of endogenous scales and how it can lead to new visualization systems for exploring knowledge dynamics can be found in Lobbé et al. (2021).

Computing the ancestors beyond the time horizon

The inter-temporal matching procedure described in "Adaptive inter-temporal matching and step phylomemetic networks" section may introduce artificial splits of phylomemetic branches due to the incompleteness of the corpus analyzed, such as the non-availability of digitized documents beyond a certain date. To mitigate this artifact and ease the interpretation, 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. This algorithm is detailed in SI C.6. and only impacts the visualization of phylomemetic networks. The reconstruction operator that takes this steps into account will be subsequently noted ${}^4\overline{\boldsymbol{\Phi}}$.

Results

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

We present here a quantitative evaluation of the new workflow $\phi = {}^4\boldsymbol{\Phi}\circ{}^3\boldsymbol{\Phi}\circ{}^2\boldsymbol{\Phi}\circ{}^1\boldsymbol{\Phi}$ compared to results obtained by Chavalarias and Cointet (2013), as well as a qualitative assessment of its ability to accurately describe the evolution of scientific fields. The perspectives offered by this new methodology for the interaction with large set of documents through visualization are further detailed in Lobbé et al. (2021) and its contribution to the analysis of history and philosophy of science is detailed in Chavalarias et al. (2021).

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

- $\Xi(x, \_)$ 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 "Discussion" section, 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 Eq. 3:

$$F_\lambda(\varphi) = \Sigma_{x\in\mathcal{L}} \frac{1}{|\mathcal{L}|} . \Sigma_{B_k\in\mathcal{B}_k^x} \frac{|B_k|}{\Sigma_{B_j\in\mathcal{B}_k^x}|B_j|} . F_\lambda(x,k) \tag{3}$$

where $\mathcal{B}_k^x = \{B_k | B_k \cap \mathcal{C}_x \neq \varnothing\}$

Quantitative and qualitative validation

We have chosen several distinct case studies to illustrate the wide range of applications of our methodology and the robustness of our results. These case studies have been delineated thanks to the following corpora (see SI D. for full details):

- ***Domain specific academic literature.***
  - $\mathcal{D}_{glyphosate}$: *Glyphostate literature*. A corpus of 16,7 k documents retrieved in the Web of Science (WoS) and PubMed.
  - $\mathcal{D}_{quantum}$: *Quantum computing literature*. A corpus of 29 k documents retrieved from the WoS.
- ***Interdisciplinary academic literature.*** $\mathcal{D}_{CNRS}$: a corpus of 6000 top-cited CNRS papers from the WoS. CNRS being an interdisciplinary research organism, this corpus is by construction highly interdisciplinary. Although it contains a limited number of docu-

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

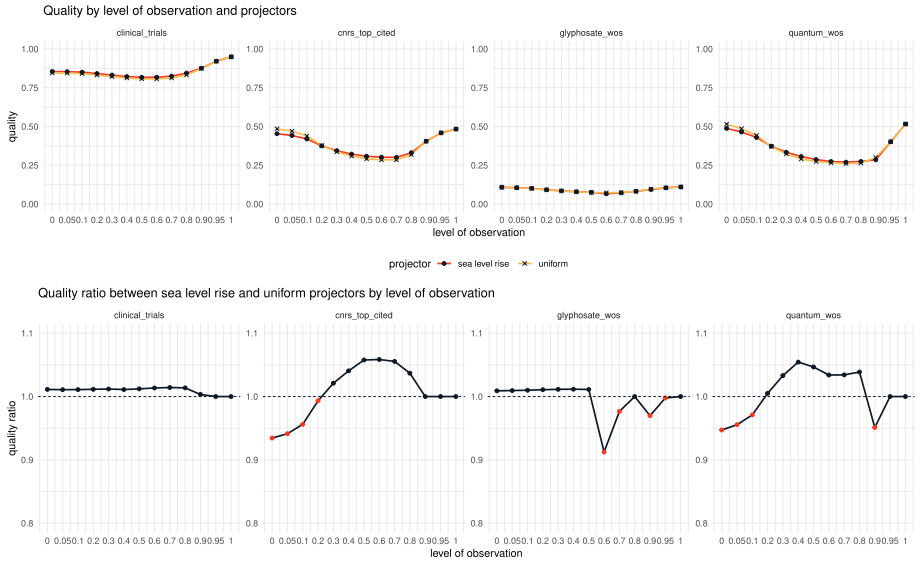

Fig. 7 Comparison between the sea-level rise and uniform projectors for four distinct corpora. Clustering method: frequent item sets (cf. Table SI 2 in SI E.)

ments in each discipline, the phylomemy reconstruction succeed in providing an overview of the main research streams and highlights the way they combine (Lobbé et al. 2021).

– ***Generic time-stamped corpora.*** Our method can reconstruct knowledge dynamics from any kind of time-stamped corpora and can process very short documents as well. As an illustration, we have applied our method to a corpus $\mathcal{D}_{CT}$ consisting in 6000 records of clinical trials arms related to Covid-19 treatments. Associated phylomemetic networks highlight the different research paths and discoveries made around the Codiv-19 outbreak, a useful knowledge for the biomed community (cf. SI G.3. for details and Lobbé et al. 2021 for the same analysis on Covid-19 vaccines).

For this comparison, we have adopted the *confidence* similarity measure[18] in ${}^{2}\boldsymbol{\Phi}$ and we use either the *maximal cliques* or the *frequent item sets* (FIS) clustering methods in ${}^{3}\boldsymbol{\Phi}$. Full details regarding the implementation and settings of the phylomemy reconstruction workflow are provided in SI C. and SI E. respectively.

Quantitative evaluation

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. The question is whether the quality of the phylomemies obtained with the sea-level rise algorithm could be significantly higher than the quality reached by

[18] 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.

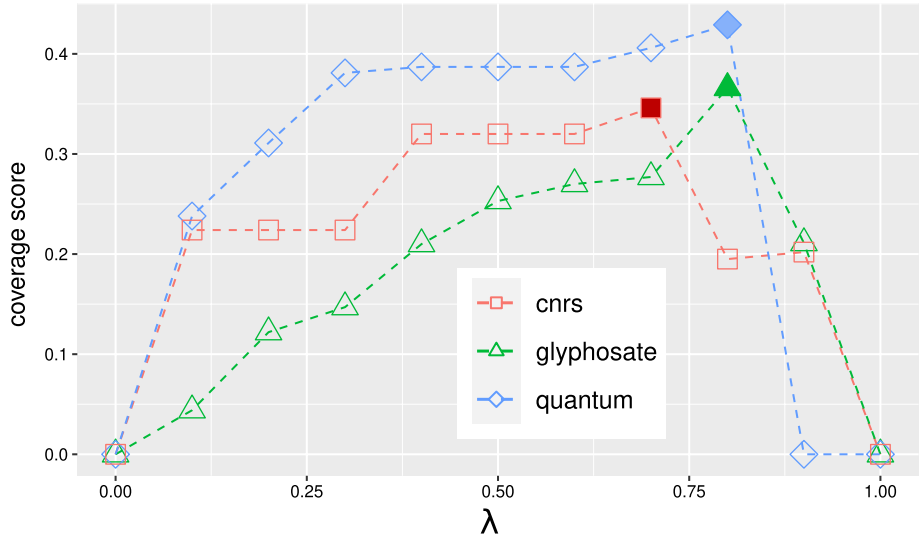

Fig. 8 Sensitivity analysis of the coverage scores for different case studies. For all case studies, the coverage score reaches its maximum at a λ value that constitutes an interesting entry point for the exploration of a phylomemy. For example, the preferred entry point for $\mathcal{D}_{glyphosate}$ is 0.8, the value chosen for plotting Fig. 3

uniform projectors. For this comparison, we will consider the family of uniform projectors $\{\pi_\delta | \delta \in \{0, 0.1, \ldots, 0.9, 1\}\}$.

As can be seen on Fig. 7, for the various case studies and for most levels of observation λ, step phylomemetic networks obtained with sea level algorithm outperform or are at least as good as the best uniform projectors. We moreover demonstrates in SI F.1. 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.

We insist on the fact that the quantitative analysis of a set of $\{\varphi_\lambda\}_\lambda$ for different F_λ scores can't be considered as a comparison between these φ_λ since the objective functions $\{F_\lambda\}_\lambda$ are different. Consequently, this type of analysis does not point to a preferred λ value for the observation of a phylomemy. The appropriate value depends on what the analyst is looking for.

We can however determine a preferred entry point for the exploration of a phylomemy. Let's consider a phylomemy where small branches have been filtered out. As the level of observation decreases, φ_λ consists of more branches of lower complexity, while more and more $\mathcal{C}^*$ clusters have no inheritance links which leads to a higher number of small branches not appearing in the network. Therefore, levels of observation close to 1 feature few branches with almost all terms being contextualized by a branch ; whereas levels of observation close to 0 describe a variety of subdomains but fail to contextualize a significant proportion of $\mathcal{L}$. We can assume that an analyst would want to start her exploration with a phylomemetic network that have both a good coverage of $\mathcal{L}$ and a good description of the sub-domains concerned by the corpora under study. We can thus define a *coverage*

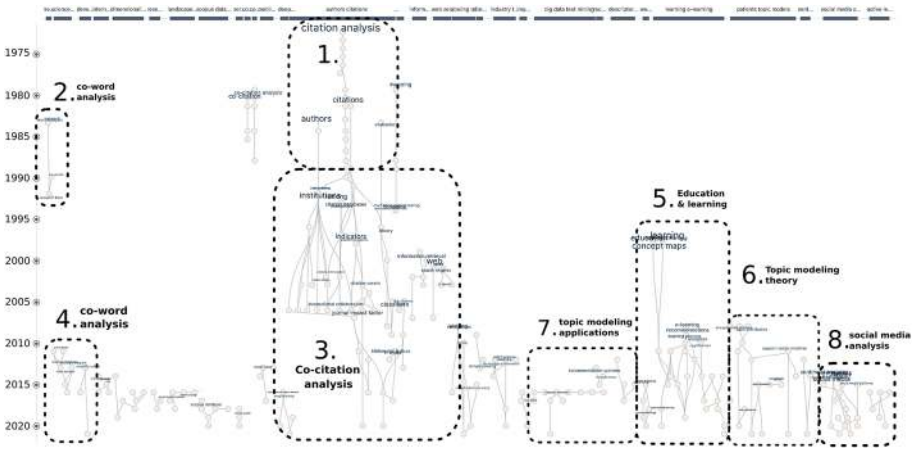

Fig. 9 Phylomemetic network of the literature related to science and knowledge mapping ($\mathcal{D}_{maps}$) at level 0.3 with branches smaller than 3 filtered out. Dashed boxes have been manually annotated. An interactive version of this phylomemetic network available at http://maps.gargantext.org/phylo/knowledge_visualization/memiescape and can be downloaded from the archive (Quentin Lobbe and Chavalarias 2021)

score for a given phylomemetic network as the geometric mean of its relative roots coverage and its normalized number of branches.[19] The maximum of this score points to a preferred λ value for the exploration of a phylomemy as illustrated by Fig. 8.

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*.

Reconstruction of the history of a research domain: the example 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. Literature on glyphosate is quite recent, most of it being digitized. The knowledge it contains is of great importance from a health and economic point of view. Moreover, there is yet no consensual synthesis of the knowledge this research as produced to far. Glyphosate is a controversial herbicide about which literature reviews and historical analyses are regularly published, some emphasizing the advantages of glyphosate or the absence of associated risks, others reviewing the risks

[19] Relative root coverage (in [0 1]) is computed relatively to the maximum and minimum proportion of roots covered by a phylomemetic network; normalized number of branches is relative to the maximum number of branches of phylomemetic networks when λ varies.

for health and environment and issues related to the the emergence of herbicide-resistant weeds.

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, 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.

To compare the picture depicted by the method of phylomemy reconstruction with that depicted by experts in the field, we have synthesized the most cited literature reviews written by glyphosate supporters and skeptics (Duke 2018; Benbrook 2016; Székács and Darvas 2012; Gillezeau et al. 2019; Martinez et al. 2018; Singh et al. 2020).

The analysis of the phylomemy of glyphosate research at different scales of observation (cf. Fig. 3 and SI G.2.) makes it possible to successfully identify the different research questions present in this synthesis, the details of their ramifications and their development. Full details of this analysis are provided in SI G.

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

The phylomemetic networks for $\lambda = 0.3$ of the knowledge dynamics corpus $\mathcal{D}_{maps}$ analyzed in "Mapping science and knowledge" section is presented on Fig. 9.

We can observe that this phylomemetic networks 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 (Terzopoulos 1985; Callon et al. 1983) 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).

Discussion

Future improvements of the sea-level rise algorithm

Red dots on Fig. 7 highlight the values of λ for which the current implementation of the sea-level rise algorithm does not overtake the uniform projectors. For these data point at

least, there are optimization margins in the way we locally increment and elect δ in this algorithm (see SI C.5). At least two factors explain this sub-optimality. First, finding the step phylomemetic network that optimizes F_λ is done in a rugged quality landscape, especially for the choice of the elevation step. It is therefore a difficult task that deserves a paper in itself. The algorithm proposed in this paper is a only a starting point an can undoubtedly be improved. Second, for computing time reasons and computational resources, we have filtered the clusters to some extent (cf. Table SI 2). This operation might have eliminated some optimal step phylomemetic networks.

As for the scalability of this algorithm, the computational complexity is more or less linear regarding the number of documents but depends heavily on the size of the list of terms upon which the phylomemy is reconstructed and on the clustering algorithm chosen to define the fields. In the example chosen in this paper, fields are defined as maximal cliques, which worst-case time complexity is $O(3^{\frac{n}{3}})$ for an n-vertex graph (Tomita et al. 2006). Hopefully, semantic networks are generally sparse so that the maximal cliques algorithm has always been tractable in reasonable time.

The computational complexity also depends on the number of clusters generated by the clustering algorithm. The computation of inheritance links scales as $O(N_c^3)$, where N_c is the number of clusters. However, we only need to compute the similarities between clusters with non empty overlap and local optimizations can be done to drastically reduce the number of interesting clusters such as keeping only the most cohesive ones.

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 Chuang et al. 2012 for a good example).

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 (1751):

> 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.

Table 1 Comparison between different approaches for the modeling of knowledge dynamics

Paper	Domain	Focus	Unit of analysis	Methods	Fields detection	Meaning
This paper	SoS	Corpora based	Text	CWA	COC	P+S
Rule et al. (2015)	SoS	Corpora based	Text	CWA	COC	P+S
Wang et al. (2014)	SoS	corpora based	Text	CWA	COC	P+S
Chavalarias and Cointet (2013)	SoS	Corpora based	Text	CWA	COC	P+S
Shahaf et al. (2013)	IR	Corpora based	Text	CWA	COC	P+S
Liao and Qian (2019)	IR	Corpora based	Documents	CNA	MGF	×
Tacchella et al. (2020)	ML	Corpora based	Codes	WE	×	P
Palmucci et al. (2019)	ML	Corpora based	Codes	WE	×	P
Jähnichen et al. (2018)	IR	Corpora based	Documents	TM	DTM (LDA)	P
Chen et al. (2017)	IR	Corpora based	Documents	TM	LDA	P
Yang et al. (2017)	Visual analytics	Corpora based	Documents	TM	HLMT	×
Wang et al. (2015)	IR	Corpora based	Documents	TM	dDTM	–
Cui et al. (2011)	IR	Corpora based	Documents	TM	HDP	×
Gohr et al. (2009)	IR/ML	Corpora based	Documents	TM	PLSA	×
Cambe et al. (2020)	SoS	Corpora based	Documents	CA	BC	×
Claveau and Gingras (2016)	SoS	Corpora based	Documents	CA	BC	S
Jo et al. (2011)	SoS	Query based	Documents	TM + CA	PLS + CA	×
Shahaf et al. (2012)	IR	Query based	Documents	IR	×	×

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)

Table 2 Comparison between different approaches for the modeling of knowledge dynamics

Paper	This paper	Liao and Qian (2019)	Shahaf et al. (2013)	Chavalarias and Cointet (2013)	Jo et al. (2011)	Jähnichen et al. (2018)	Wang et al. (2015)	Gohr et al. (2009)	Chen et al. (2017)
Publication year	2021	2019	2013	2013	2011	2018	2015	2009	2017
Method	CWA	CNA	CWA	CWA	TM+CA	TM	TM	TM	TM
Allow re-emerging topics ?	✓	✓	✓	✓	✓	×	×	×	×
Allow split/merge events	✓	✓	✓	✓	✓	×	×	×	✓
Evolving topics?	✓	✓	✓	✓	✓	✓	✓	✓	✓
Unconstrained # of topics	✓	✓	✓	✓	✓	×	×	×	×
Objective function	✓	✓	✓	×	×	✓	✓	✓	×
Multi-level	✓	×	⋯	⋯	⋯	×	×	×	×
Multi-scale	✓	✓	✓	✓	×	✓	✓	×	×
Advanced text-mining?	✓	×	×	✓	×	×	×	✓	×
Work on unstructured text?	✓	×	✓	✓	×	✓	✓	×	✓
Work on short texts?	✓	⋯	✓	✓	?	?	×	×	×
Work on less than 10 k doc ?	✓	✓	✓	✓	?	?	?	?	?
Global science maps?	✓	×	✓	✓	✓	✓	✓	✓	✓
External validation	✓	✓	✓	✓	×	✓	×	×	×
Internal/quantit. evaluation	✓	✓	✓	✓	×	✓	✓	×	×
Integrate users preferences?	✓	⋯	⋯	×	×	×	×	×	×
Advanced visualizations	✓	×	✓	✓	✓	×	×	✓	✓
Software reproducibility	✓	×	?	?	×	?	×	×	×
Open source	✓	×	?	×	×	?	×	×	×

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)

Table 3 Comparison between different approaches for the modeling of knowledge dynamics

Paper	Claveau and Gingras (2016)	Rule et al. (2015)	Wang et al. (2014)	Cui et al. (2011)	Cambe et al. (2020)	Tacchella et al. (2020)	Palmucci et al. (2019)	Shahaf et al. (2012)	Yang et al. (2017)
Publication year	2016	2015	2014	2011	2020	2020	2019	2012	2017
Method	CA	CWA	CWA	TM	CA	WE	WE	IR	TM
Allow re-emerging topics?	×	×	×	×	×	–	–	–	✓
Allow split/merge events	✓	✓	✓	✓	✓	–	–	✓	×
Evolving topics?	✓	✓	✓	✓	✓	–	–	–	×
Unconstrained # of topics	✓	✓	✓	✓	✓	–	–	–	✓
Objective function	×	×	×	×	✓	–	–	✓	✓
Multi-level	×	×	×	×	×	×	–	···	×
Multi-scale	×	···	✓	✓	✓	×	–	✓	✓
Advanced text-mining?	×	✓	×	×	×	×	×	×	?
Work on unstructured text ?	×	✓	✓	✓	×	···	···	✓	✓
Work on short texts?	n/a	✓	✓	×	✓	✓	✓	×	?
Work on less than 10 k doc?	✓	✓	✓	?	✓	×	×	✓	?
Global science maps?	✓	✓	✓	✓	✓	✓	✓	×	✓
External validation	×	✓	✓	✓	✓	✓	✓	✓	✓
Internal/quantit. evaluation	×	×	×	×	✓	✓	✓	✓	✓
Integrate users preferences?	×	×	/	···	×	×	×	✓	×
Advanced visualizations	✓	✓	✓	✓	✓	×	×	✓	✓
Software reproducibility	×	✓	?	?	✓	×	×	?	?
Open source	×	?	?	?	✓	×	×	?	?

Legend: same as Table 2

The objective function described by Eq. 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. A preliminary study of alternative workflows is presented in SI F., where alternative objective functions, inter-temporal matching function and alternative clustering algorithms have been compared on the different case studies.

Our main conclusion is that comparing different reconstruction workflows is not only about improving quality with respect to F_λ. It also involves comparing the respective contributions of each option to the characterization, at different levels and scales, of the case studied. Going back to D'Alembert's analogy mentioned in "Introduction" section, 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.

Comparison with previous works

In order to compare the proposed method with previous works mentioned in the state-of-the-art, we have identified a set of key properties fulfilled by our own contribution—listed in SI B.2.—and have reviewed the way they appear (or not) in a sub-collection of representative past papers (cf. Tables 1, 2 and 3). We limit here the discussion to papers that can handle unstructured text.

Most of the papers reviewed do not allow for re-emerging topics: they only perform inter-temporal matching between two consecutive periods of time. This is a severe limitation since there is no reason to assume that the parents of a field of knowledge belong to the immediate previous period. Moreover, this assumption makes these methods very sensitive to the initial slicing of the time periods. Among the research papers that consider re-emerging topics, Chen et al. (2017) does not consider split or merge events nor evolving topics. Consequently, it can't analyze the structure of knowledge dynamics or the evolution of knowledge fields. Liao and Qian (2019) adopts an information retrieval perspective where the entry point is a paper or a set of papers and the building block of the maps are single articles. Consequently, this approach does not aim at producing global science maps but rather chains of specific papers that contextualize the initial query. This approach is also not scalable with respect to the number of papers.

The works that are most related to our approach are thus Shahaf et al. (2013), Chavalarias and Cointet (2013) and Jo et al. (2011). Jo et al. (2011) produces objects that are similar to our phylomemetic networks from a topic modeling perspective. However, the authors do not give a meaning to the global inter-temporal matching procedure. Since there is no objective function to determine the choice of inter-temporal threshold, they struggle to interpret threshold values and leave the reader with only the observation that a threshold that is "effective in revealing the evolution structure of dense areas [...] does not discover the structures for sparse area." Without further precision, the initial goal could not be accomplished. The lack of objective function is also what prevents (Chavalarias and Cointet 2013) to give full insight into the topology of knowledge dynamics.

Last, Shahaf et al. (2013) did adopt the same building blocks than Chavalarias and Cointet (2013) and this paper. Moreover, they have several objective functions to filter out the interesting knowledge dynamics branches. However, these branches are built from a global clustering algorithm on all the knowledge fields independently of their periods.

Consequently, this approach cannot entangle time-scales (in terms of how far in time you search for predecessors or successors of a field) and inter-temporal matching strength (what is the weakest eligible association). As a result, while they do have a notion of scale, their method cannot convey the notion of level of observation.

Embodied cognition and users' preferences

Although it is common practice to take users' preferences into account in information retrieval tasks (cf. Druck et al. 2008; Shahaf et al. 2012 for examples), it is much less common in the literature on science and knowledge mapping (left part of Fig. 1). 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 the aim of part of this literature 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 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 collective of users. These constraints lead us to make the distinction between two classes of adaptation to the user's preferences:[20]

- The choices of a level and a scale of observation allow users to agree on the intrinsic and extrinsic complexity of the representation they want to share,
- 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 interaction of users with different answers to the same question $\mathcal{Q}(x)$. 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 F.1., where an alternative formulation for Ξ has been tested.

If we imagine that phylomemy reconstruction can be used by a community of scholars to either retrieve documents or collectively assess the shapes and properties of a research landscape, then Ψ and Ξ should be estimated and revised according to the collective behavior of scholars. Phylomemetic networks can then play, at the collective level, a role

[20] 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.

analogous to individual mental representations, and these collective representations can become true tools for the coordination among scholars. It is important to note that in such settings, these representations would be co-constructed through the users' interactions with a digital environment. In that case, we cannot consider the representations provided by the system as external to its users anymore.

Phenomenological reconstruction, as formalized in this paper by the chain $\mathcal{O} \longmapsto \mathcal{R} \longmapsto \mathcal{V}$, is consequently 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 thus have a circular dependency between the preferences of the members of a collective that interact with a phenomenological reconstruction and the parameters of this same reconstruction. This circular dependency can be related to Francisco Varela's conception of cognition (Varela 1979), 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 (Varela et al. 2000). Collective cognition emerges from morphogenetic and path-dependent processes during our interactions with our environment. By simplifying reality through 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.

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 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,
- 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,
- 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 temporal semantic networks as a natural output of the process of phylomemy reconstruction,
- we have integrated users' preferences into our framework by providing an interaction model and contextualizing the different elements of our reconstruction workflow in the theoretical framework of Varela's embodied cognition,
- by applying our method to the state-of-the-art of this paper, we have illustrated how it could be applied to specify the positioning of scientific articles.

The diversity of our case studies demonstrates that we can address with the same methodology a wide variety of textual contents, from big data to small data, and from short texts to full texts, where other approaches are more focused on specific types of data. This paves the way to achieving D'Alembert's dream for a large range of knowledge production arenas via a unified methodology (*e.g.* academic literature, patents, news, [micro-]blogs, etc.).

The formalisation of the notion of levels and scales allows the user to navigate and interact intuitively with the knowledge dynamics. Implemented in a dedicated visualization software (Lobbé et al. 2021) phylomemy reconstruction thus offers the philosopher "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" (d'Alembert 1751).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-021-04186-5.

Acknowledgements We warmly thank Bruno Gaume for his fruitful comments on our work. Funding was provided by Agence Nationale de la Recherche (Grant Nos. ANR-16-CE38-0002-01, ANR-11-IDEX-0005-02 and ANR Flag-ERA JTC 2016).

Funding 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 (ANR-11-IDEX-0005-02) Project and the EU FuturICT 2.0 Project (ANR Flag-ERA JTC 2016).

Data transparency and replication All data and code used for this paper have been published in open access for full reproducibility: Supporting information gives details about some aspects of this paper. Data and supplementary material are available on the archive David Chavalarias et al. (2021). A dedicated software for interactive visualization is available on the archive Quentin Lobbe et al. (2021) and presented in Lobbé et al. (2021). The code developed 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).

Declarations

Conflict of interest There is no conflict of interest in this submission and all authors have approved the final version.

References

Abell, S. K., & Lederman, N. G. (Eds.). (2007). *Handbook of research on science education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI conference on weblogs and social media*.

Benbrook, C. M. (2016). Trends in glyphosate herbicide use in the United States and globally. *Environmental Sciences Europe, 28*(1), 3. https://doi.org/10.1186/s12302-016-0070-0.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Bourgine, P., Brodu, N., Deffuant, G., Kapoula, Z., Müller, J. P., & Peyreiras, N. (2009). Formal epistemology, experimentation, machine learning. In *HAL archives ouvertes*. https://hal.archives-ouvertes.fr/hal-00392486, pp. 10–14. Chavalarias et al.

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(12), 2389–2404. https://doi.org/10.1002/asi.21419. http://doi.wiley.com/10.1002/asi.21419.

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*(4), 233–251. 10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I.

Börner, K. (2010). *Atlas of science: Visualizing what we know*. Cambridge, MA: The MIT Press.

Börner, K. (2015). *Atlas of knowledge: Anyone can map*. Cambridge, MA: The MIT Press.

Börner, K., Chen, C. M., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37,* 179–255. https://doi.org/10.1002/aris.1440370106.

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*(2), 191–235. https://doi.org/10.1177/053901883022002003.

Callon, M., Rip, A., & Law, J. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. Springer.

Cambe, J., Grauwin, S., Flandrin, P., & Jensen, P. (2020). Exploring and comparing temporal clustering methods. arXiv:2012.01287 [physics].

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.*https://doi.org/10.1098/rstb.2019.0329. *https://royalsocietypublishing.org/doi/10.1098/rstb.2018.0348.*

Chavalarias, D., & Cointet, J. P. (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. *Scientometrics,75*(1), 37–50. https://doi.org/10.1007/s11192-007-1825-6. http://link.springer.com/10.1007/s11192-007-1825-6.

Chavalarias, D., & Cointet, J. P. (2013). Phylomemetic patterns in science evolution–the rise and fall of scientific fields. *PloS one,8*(2), e54847. http://dx.plos.org/10.1371/journal.pone.0054847.

Chavalarias, D., Huneman, P., & Racovski, T. (2021). Using phylomemies to investigate the dynamics of science. In *The dynamics of science: Computational frontiers in history and philosophy of science*. Pittsburgh University Press, Pittsburgh.

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*(4), 1175–1189. https://doi.org/10.1016/j.joi.2017.10.003. http://www.sciencedirect.com/science/article/pii/S1751157717300536.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology,57*(3), 359–377. https://doi.org/10.1002/asi.20317. http://doi.wiley.com/10.1002/asi.20317.

Chen, C. (2017). Science mapping: A systematic review of the literature. *Journal of Data and Information Science,2*(2), 1–40. https://doi.org/10.1515/jdis-2017-0006. https://content.sciendo.com/view/journals/jdis/2/2/article-p1.xml. Publisher: Sciendo Section: Journal of Data and Information Science.

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*, CHI '12, pp. 443–452. Association for Computing Machinery, Austin, Texas, USA. https://doi.org/10.1145/2207676.2207738.

Claveau, F., & Gingras, Y. (2016). Macrodynamics of economics: A bibliometric history. *History of Political Economy, 48*(4), 551–592.

Cointet, J. P., & Chavalarias, D. (2008). Multi-level Science mapping with asymmetric co-occurrence analysis: Methodology and case study. Networks and Heterogeneous Media pp. 267–276.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., & Tong, X. (2011). TextFlow: Towards better understanding of evolving topics in text. In *IEEE transactions on visualization and computer graphics*, Vol. 17, pp. 2412–2421. IEEE Educational Activities Department. https://doi.org/10.1109/TVCG.2011.239.

David, C., Quentin, L., Delanoë, A. (2021). Replication data for: Draw me science—multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. https://doi.org/10.7910/DVN/SBH3EI, Harvard Dataverse.

Dias, G., Mukelov, R., & Cleuziou, G. (2008). Mapping general-specific noun relationships to wordnet hypernym/hyponym relations. In *Knowledge engineering: Practice and patterns*, pp. 198–212. Springer.

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.

Duke, S. O. (2018). The history and current status of glyphosate. *Pest Management Science, 74*(5), 1027–1034. https://doi.org/10.1002/ps.4652.

dAlembert, J. l. R. (1751). Discours préliminaire des Éditeurs. In J. l. R. d'Alembert, & D. Diderot (Eds.), *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers*, Vol. Tome 1, pp. i–xlv.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479.

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*(1), 2. https://doi.org/10.1186/s12940-018-0435-5.

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, proceedings*, pp. 859–870. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972795.74.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE, 9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679.

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.

Jähnichen, P., Wenzel, F., Kloft, M., & Mandt, S. (2018). Scalable Generalized Dynamic Topic Models. arXiv:1803.07868 [cs, stat].

Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10. https://doi.org/10.1002/asi.5090140103.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*(5), 604–632.

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*(6), 574–582. https://doi.org/10.1205/cherd.05093.

Liao, D. P., & Qian, Y. T. (2019). Paper evolution graph: Multi-view structural retrieval for academic literature. *Frontiers of Information Technology & Electronic Engineering, 20*(2), 187–205. https://doi.org/10.1631/FITEE.1700105.

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.

Lobbé, Q., Delanoë, A., & Chavalarias, D. (2021). Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. Information Visualization p. 14738716211044829. https://doi.org/10.1177/14738716211044829. Publisher: SAGE Publications.

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*(1), 2. https://doi.org/10.1186/s12302-018-0131-7.

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*.

Palla, G., Barabási, A. L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664–667. https://doi.org/10.1038/nature05670.

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].

Quentin, L., Delanoë, A., & Chavalarias, D. (2021). Replication data: Exploring, browsing and interacting with multi-scale structures of knowledge. https://doi.org/10.7910/DVN/WLI9B5, Harvard Dataverse.

Rossetti, G., & Cazabet, R. (2018). Community discovery in dynamic networks: A survey. *ACM Computing Surveys, 51*(2), 1–37. https://doi.org/10.1145/3172867. arXiv:1707.03186.

Roth, C., & Cointet, J. P. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16–29. https://doi.org/10.1016/j.socnet.2009.04.005.

Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, pp. 1790–2014. Proceedings of the National Academy of Sciences p. 201512221. https://doi.org/10.1073/pnas.1512221112.

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.

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*, pp. 1097–1105. ACM.

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*(3), 663–702. https://doi.org/10.1007/s10311-020-00969-z.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265–269.
Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics, 38*(2), 275–293.
Székács, A., & Darvas, B. (2012). Forty years with glyphosate. In M. N. A. E. G. Hasaneen (Ed.), *Herbicides-properties, synthesis and control of weeds* (pp. 247–284). Rijeka: IntechOpen.
Tacchella, A., Napoletano, A., & Pietronero, L. (2020). The language of innovation. *PLoS ONE, 15*(4), e0230107. https://doi.org/10.1371/journal.pone.0230107.
Terzopoulos, D. (1985). Co-occurrence analysis of speech waveforms. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 33*(1), 5–30.
Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science, 363*(1), 28–42. https://doi.org/10.1016/j.tcs.2006.06.015.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95. https://doi.org/10.1038/s41586-019-1335-8.
Varela, F. J. (1979). Principles of biological autonomy. The North Holland series in general systems research; 2. North Holland, New York.
Varela, F. J., Thompson, E., & Rosch, E. (2000). *The embodied mind: Cognitive science and human experience* (8th ed.). Cambridge, MA: MIT Press.
Wang, C., Blei, D., & Heckerman, D. (2015) Continuous time dynamic topic models. arXiv:1206.3298 [cs, stat].
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*, p. 448. ACM Press, San Diego, California, USA. https://doi.org/10.1145/2020408.2020480. http://dl.acm.org/citation.cfm?doid=2020408.2020480.
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*(2), 1253–1271. https://doi.org/10.1007/s11192-014-1347-y.
Weeds, J., & Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics, 31*(4), 439–475. https://doi.org/10.1162/089120105775299122.
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.
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*(4), 327–355.
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*(1), 40–47.