



# Unpacking research lock-in through a diachronic analysis of topic cluster trajectories in scholarly publications

Matteo Lascialfari<sup>1</sup> · Marie-Benoît Magrini<sup>1</sup> · Guillaume Cabanac<sup>2</sup>

Received: 14 July 2021 / Accepted: 2 September 2022 / Published online: 27 September 2022  
© Akadémiai Kiadó, Budapest, Hungary 2022

## Abstract

Lock-in and path-dependency are well-known concepts in economics dealing with unbalanced development of alternative options. Lock-in was studied in various sectors, considering production or consumption sides. Lock-in in academic research went little addressed. Yet, science develops through knowledge accumulation and cross-fertilisation of research topics, that could lead to similar phenomena when some topics do not sufficiently benefit from accumulation mechanisms, reducing innovation opportunities from the concerned field consequently. We introduce an original method to explore these phenomena by comparing topic trajectories in research fields according to strong or weak accumulative processes over time. We combine the concepts of ‘niche’ and ‘mainstream’ from transition studies with scientometric tools to revisit Callon’s strategic diagram with a diachronic perspective of topic clusters over time. Considering the trajectories of semantic clusters, derived from titles and authors’ keywords extracted from scholarly publications in the Web of Science, we applied our method to two competing research fields in food sciences and technology related to pulses and soya over the last 60 years worldwide. These highly interesting species for the sustainability of agrifood systems experienced unbalanced development and thus is under-debated. Our analysis confirms that food research for soya was more dynamic than for pulses: soya topic clusters revealed a stronger accumulative research path by cumulating mainstream positions while pulses research did not meet the same success. This attempt to unpack research lock-in for evaluating the competition dynamics of scientific fields over time calls for future works, by strengthening the method and testing it on other research fields.

---

✉ Marie-Benoît Magrini  
Marie-benoit.magrini@inrae.fr

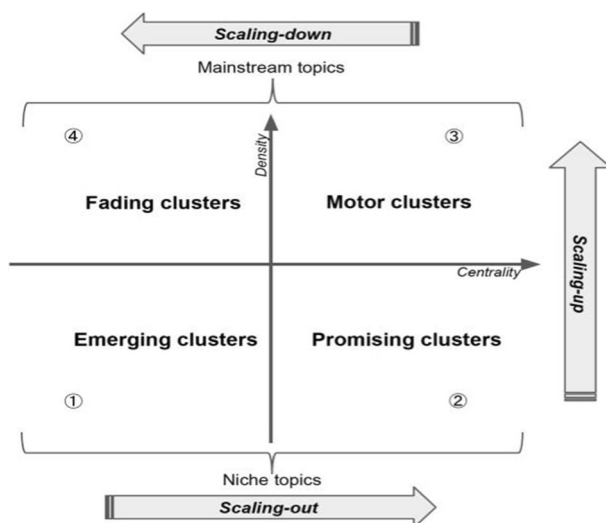
Matteo Lascialfari  
mtt.lascialfari@gmail.com

Guillaume Cabanac  
guillaume.cabanac@univ-tlse3.fr

<sup>1</sup> AGIR, Université de Toulouse, INRAE, Castanet-Tolosan, France

<sup>2</sup> IRIT, Université de Toulouse, CNRS, Toulouse, France

## Graphical abstract



**Keywords** Bibliometric data · Path-dependency · Natural language processing · Science mapping · Sustainability concern

## Introduction

Lock-in is a key problem that hampers sustainability transition. We found no work analysing how and why science could also suffer from lock-in. Since the seminal works of Arthur (1994), the characteristics of lock-in are mainly depicted by economists as technological lock-in<sup>1</sup> through production and consumption markets and less in research per se. For instance, in the recent review of Conti et al. (2021), the authors identified 122 articles dealing with lock-in or path-dependency in agrifood sector. Those papers identified various mechanisms explaining lock-in as the inability of society or subgroups of actors to move toward alternatives, although these may bring improvements for society. None of these studies, however, questioned lock-in in academic research regarding the balanced development of various fields of knowledge from which alternatives could rely for favouring innovation and direction turns in society. It is indeed a difficult question but we believe scientometrics should strive to identify such situations in order to inform science and innovation policy, particularly when it relates to a major issue for sustainability.

<sup>1</sup> Brian Arthur defined technological lock-in as a situation where two technologies compete and over time one technology turns to be the dominant one on the market (through increasing return adoption mechanisms) unlike the other one could be more beneficial to society. Lock-in is the consequence of path-dependency. The path-dependency concept highlights the dependence to past/previous choices or investment while the lock-in concept highlights the inability of alternatives to develop or to compete “as equal/on par” as one technology turned to be the dominant one by benefiting more quickly of accumulation effects (ie. path-dependency).

Even though scientific knowledge co-evolves with production and consumption dynamics (e.g. Dosi & Nelson, 2010) and several authors recognised that scientific knowledge is also characterised by cumulative and path-dependent mechanisms (e.g. Xu et al., 2020; Hu & Rousseau, 2018; Heimeriks & Boschma, 2014), very few authors have questioned lock-in in science. Peacock (2009) suggested an analogy to compare the works of Kuhn (1970) on science building by an accumulation process with those of Arthur on technological lock-in. However unpacking research lock-in to reveal disequilibria in terms of research investment or knowledge accumulation between research fields still remains challenging. By knowledge accumulation we consider both the fact that the quantity of knowledge is increasing and leads to more and more structuration and development of new topics. As we failed to find any research that illustrates methods for measuring or characterising with data, and particularly with bibliometric data, such a process in science, we developed a method for analysing this path-dependency that could result in a lock-in phenomena in science in case of two competing fields of research. The illustration and understanding of lock-in in science will be informative for both policy makers and researchers.

Overall, scholarly publications remain the main data used to characterise the evolution of sciences (Chavalarias & Cointet, 2008, 2013; and for instance, Rafols et al., 2014 on big pharma or Epicoco et al. (2014) on green chemistry); even if works have been developed using other information such as patents (e.g. Xu et al., 2020; Sorenson & Fleming, 2004; Balconi et al., 2004) or the grey literature (e.g. Adams, 2016). We chose to work on scholarly publications and on two competing fields in agrofood sector, as a case study, because of their increasing importance in sustainability transition issues. The field we selected deals with the development of grain-legumes; acknowledgement of the importance of these crops' agro-environmental and nutritional properties is definitely on the increase (Hallström et al., 2015; Jallinoja et al., 2016; Peoples et al., 2019; Willett et al., 2019; Weindl et al., 2020; Semba et al., 2021; Cusworth et al., 2021) since the UN's International Year of Pulses in 2016.

Indeed, many studies recommend a major increase in regular consumption of pulses and soya to achieve healthy effects in diets (Abdullah et al., 2017; Havemeier et al., 2017), which currently represents an average of 7 kg a year per person in the world (FAOstats) and 4 kg a year per person in Europe (Eurostats). Poux and Aubert (2018) proposed to reach an individual intake of 11 kg a year in Europe, whereas Willett et al. (2019) advanced 18 kg a year per person for a universal healthy diet. Such a rise in consumption would consequently foster higher legume cultivation, favouring both sustainable and healthy diets. However legumes encompass an unbalanced development as a notable imbalance exists between soya and pulses with regards to their production and consumption. As a result of lock-in mechanisms, soya has become the dominant legume crop for feed and food in plant protein markets, whilst pulse production and consumption dramatically declined after the 1960s (Foyer et al., 2018; Magrini et al., 2016, 2018). The accumulation path on soya is still strong: for instance, 90% of new plant-protein foodstuffs launched in the first decade of 2000s were based on soya and wheat (Guéguen et al., 2016). Currently, whatever the global region considered, food product launched on markets with soya are more numerous than the ones with pulses (Magrini et al., 2022). Compared to soya, few food innovations are based on pulses and they are mostly characterised by incremental innovation (Lascialfari et al., 2019). Therefore, the sustainable development of both soya and pulses is challenged by this path-dependency in markets. This consideration raises also the issue about the path-dependency phenomenon regarding academic research, that constitutes the knowledge-base from which innovations could emerge, particularly regarding the field of food sciences and technology.

Recent papers suggest that this lack of innovation on pulses also relates to lower research investment. Manners and van Etten (2018) and Sonnino (2016) show that pulses benefit from less research activity than major crops such as soya. Magrini et al. (2019) found a distinct imbalance in research output on a global scale among grain-legumes: over the last decades the main crop studied was soya, and scientific publications mentioning soya alone are twice more frequent than all those mentioning any pulse species. This is clearly an unsustainable situation as scientific research is a key resource in fostering pulse development (Pinto et al., 2016). But why should such a lock-in prevail amongst pulses? Our research aims both at: (i) analysing the research output on these two competing fields—soya versus pulses—and reveal any lock-in situation. Answering this question would foster sound science policies for cracking lock-in; (ii) developing a method to answer this question that could be reproduced for other comparisons of scientific competing fields, and contribute to untangle the challenging and complex issue of lock-in phenomena in science per se.

We analysed scientific knowledge dynamics through text-mining and scientometric tools. Scientometrics, namely the quantitative study of science and innovation, assesses scientific research and its ability to answer societal needs (Chavalarias & Cointet, 2008; Ciarli & Ràfols, 2019; Glänzel et al., 2019). Relying on the seminal works of (Callon et al., 1991) on the semantic analysis of research output (publications), various studies analysed the conceptual structure of a scientific discipline. For instance, scholars identify the state-of-the-art of a scientific field in order to approach what constitutes the meso-structure of a field (Chavalarias & Cointet, 2009), to identify paradigm shifts, knowledge gaps and emerging topics, to generate hypothesis (Stegmann & Grohmann, 2003), and to forecast future developments (Courtial et al., 1993; Daim et al., 2006; Lee & Su, 2011; Moed et al., 2004; Prabhakaran et al., 2018). In order to investigate specifically lock-in analysis, we developed an original approach in analysing the dynamics of a scientific field. We enhanced Callon's strategic diagram with a longitudinal perspective, through co-word analysis, clustering and phylomemy. In other words, to first tackle this issue of lock-in analysis in science per se, we opted for this longitudinal analysis of semantic clusters to depict the path dynamics in research: which potential lock-in can be revealed according to those trajectories?

Given the strong lock-in observed in food markets concerning legumes and even more pulses, our study focuses on “Food Sciences and Technology” (FST). This research field is considered as a scientific domain particularly strategic in sustainability agrifood transition, but in which soya and pulses are in strong competition. FST encompasses various specialisations such as food microbiology, engineering and chemistry, food-making, processing, preservation and packaging, as well as nutrition, food-related allergic issues, and the psychological and sensorial analysis of food (Borsi & Schubert, 2011; Hotchkiss & Potter, 1998). Based on expert assessments (Magrini et al., 2019) we considered these specialisations as four main fields of interest—Processing, Nutrition, Allergy, and Sensory Analysis—to build our bibliographic corpus retrieved from the Web of Science Core Collection. This corpus is composed of nearly 40,000 records of scientific papers published between 1956 and 2017 on soya or pulses in the FST field on a global scale. Applying our novel approach—based on a longitudinal perspective of Callon diagram—by text-mining the records' titles and authors' keywords over time, we revealed unbalanced trajectories of FST research on soya and pulses. Pulses are in a relatively locked-in position compared to soya research which is more dynamic: soya's semantic clusters are generally more developed as clusters increased in density and centrality over time. In short, topics on soya were in mainstream positions more often than pulses, suggesting a stronger cumulative research

path while research topics on pulses did not show such effects. Based on those results and on our methodological approach, our work offers a first contribution to the analysis of path-dependency and lock-in in research that calls for subsequent exploration and measurement of such phenomena in research.

The “[Analysing scientific knowledge dynamics through text mining: a renewed approach](#)” Section presents the theoretical foundations on which we developed our approach. The “[Materials and methods](#)” Section describes the data and the method used to text-mine the FST literature on soya and pulses. The “[Results and Discussion](#)” Section presents and discusses our results, emphasizing the lock-in situation of soya vs. pulses and opens research paths towards analysing lock-in in science with bibliometric data.

## **Analysing scientific knowledge dynamics through text mining: a renewed approach**

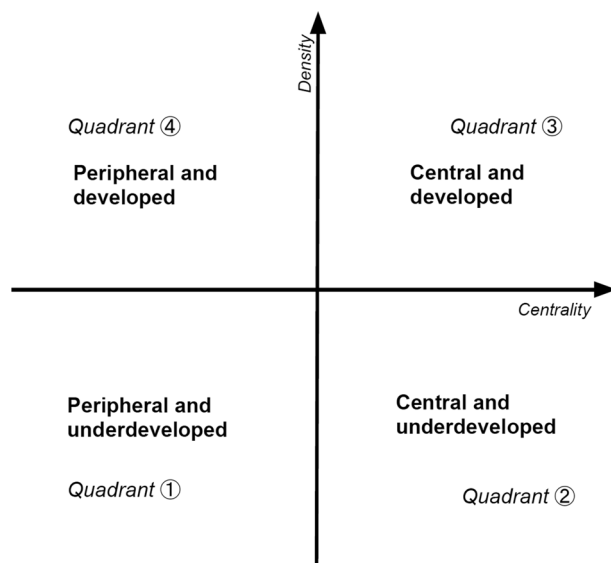
Research policies require an accurate understanding of scientific knowledge dynamics. However, tracking the evolution of a research field is challenging. Text-mining the scientific literature is broadly used when analysing scientific research evolution. The co-word approach that has been developed since the seminal works of Callon et al. (1983, 1986) draws attention to the semantic content of scientific publications to explore research trends and reveal scientific knowledge dynamics. This approach relies on text-mining relevant terms (words) within publications to map the conceptual structure of a given research field or issue (Coulter et al., 1998; Leydesdorff & Welbers, 2011). By measuring co-occurrences of terms in publications (most often for titles, authors’ keywords or abstracts) we can quantify the strength of their links (Bailón-Moreno et al., 2006), thus revealing the structure of science (Chavalarias & Cointet, 2009). Sets of terms that are strongly linked are regrouped into semantic clusters to map the main research topics of a field. This co-word analysis represents the semantic network of a scientific field, the analysis of which was progressively improved with network metrics (Jackson, 2010; Wasserman & Faust, 1994), data visualization (Drieger, 2013) and evolution analysis such as the hierarchical structure evolution approach (Qian et al., 2020).

Based on co-word analysis, we propose an original text-mining approach of scientific papers to reveal scientific knowledge dynamics over decades and address the issue of lock-in in research. Our approach is built on the Strategic Diagram—a visualisation introduced by Callon et al. (1991)—that we combined with the transition studies concepts of *niche* and *mainstream* to characterise the under-analysed semantic clusters. Finally, through inter-temporal topic-matching we shaped a diachronic analysis of research topics’ trajectories of evolution in order to reveal the main dynamics of the evolution of any scientific field.

### **The Strategic Diagram for semantic cluster analysis**

To reveal the knowledge dynamics of a scientific field, Callon et al. (1991) combined co-word analysis and clustering in a *Strategic Diagram* (SD) (Fig. 1). A SD results from the following steps: (i) words conveying meaning (stop-words having been removed) are identified and standardised; (ii) the semantic network of selected terms is built via co-word analysis; (iii) cluster detection algorithms reveal strongly related terms, representing the research topics of the field investigated; (iv) each semantic cluster is positioned in the SD

**Fig. 1** The strategic diagram in Callon et al. (1991) (adapted by the authors)



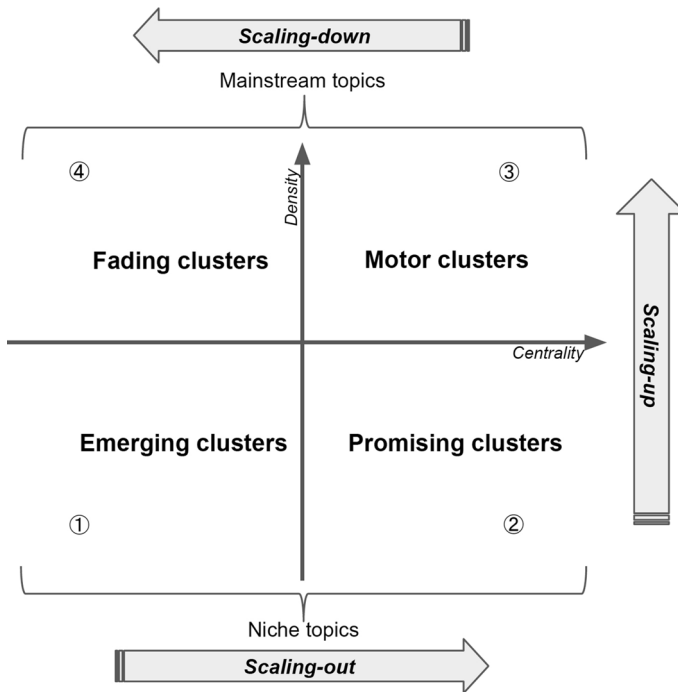
according to its degree of centrality and density (e.g. Cobo et al., 2011; Courtial et al., 1993; Glänzel et al., 2019; Yang et al., 2012).

The *density* index measures the strength of the edges that tie together each set of terms constituting a cluster, providing a representation of its internal development and capacity to maintain itself over time (Callon et al., 1991). The *centrality* index measures the proximity of a cluster to the others, thereby showing the status of a topic within the scientific field. Therefore, a cluster's location in the SD reveals its strategic position in a scientific field according to its internal development and importance for the field considered. Callon defined four quadrants depending on their position in the diagram that we consider as: Quadrant ① for research topics neither structured nor of interest to the scientific field; Quadrant ② for the under-developed but increasingly central and potentially promising clusters; Quadrant ③ for the motor subjects of the field, both dense and of central interest; Quadrant ④ for clusters well developed but of decreasing interest as they are more peripheral. Callon advanced that such positions could reveal the temporal dynamics of semantic clusters as the evolution of semantic networks is driven by two inter-related mechanisms: the reorganization of the interconnections between clusters and the internal redefinition of clusters (term changes). As a result, central clusters can disappear whilst other topics gain interest.

Despite Callon's insights, few studies have attempted to perform a longitudinal analysis on a scientific field through SD (Cahlik, 2000; Coulter et al., 1998), that is to map the life cycle of semantic clusters over time. Indeed, most studies using co-word analysis deal with the identification of the science snapshots of a field, but not its time evolution per se. This is all the more challenging as the number of publications increase every year.

### From the Strategic Diagram to a diagram revealing semantic cluster dynamics over time

Considering the approaches developed in *Transition Studies* dealing with the increasing structuration of networks, we assume that the development of a new research topic within a

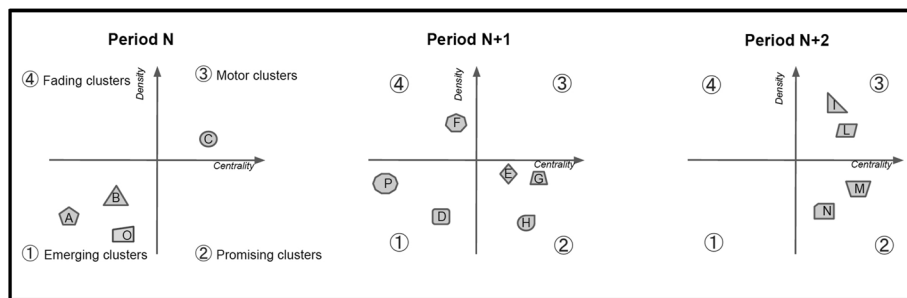


**Fig. 2** Strategic Diagram revisited through niche and mainstream concepts (adapted from Callon et al., 1991). Clusters on the lower side are considered as niche topics because of their low density, while clusters on the upper side are considered as mainstream topics because of their higher density

scientific field follows a pathway from a *niche* cluster to a *mainstream* cluster according to its degree of density. Indeed, in the Multi-Level-Perspective (MLP) approach, the development of a niche network (i.e., cluster) is sustained by an increasingly dense social network and set of concepts (Geels, 2004). Using an analogy, we can consider that more central topics represent the mainstream (or incumbent) scientific sub-regime,<sup>2</sup> whilst under-developed topics concern niches. Based on the SD, we infer that novel research topics characterised by a low density of concepts are niche clusters. These can develop further and become mainstream clusters with higher degrees of density or not, thus remaining niches or disappearing completely over time.

To take this further, we considered that research topics' dynamics are driven both by patterns of *centrality* and *density* variations through a life cycle that we characterise as follows (Fig. 2): (i) *scaling-out* concerns the increase of centrality, of external connectivity, when words/terms of a research topic are increasingly adopted by a wider range of scientists and connected to other topics; (ii) *scaling-up* when cluster words are increasingly and

<sup>2</sup> The analysis of a scientific regime as defined in transition studies is beyond the scope of this study: considering the entire scientific regime requires to consider both the topics of knowledge and the actors contributing to shape them. In particular, this extended analysis would require to analyse the socio-semantic networks to understand which social networks are at the origin of new seamless knowledge topics that could be developed by connection with mainstream social networks, and then becoming new core topics that could serve a renewal of knowledge for sustainability issues.



**Fig. 3** The Synchronic Strategic Diagram perspective over the period  $[N, N+2]$  (authors). the Synchronic Strategic Diagram is an overview of SDs mapped over different periods. This figure illustrates various semantic clusters (represented by a capital letter) appearing over the periods that differ from those mapped over the whole period

strongly tied together, and the topic gains in complexity; iii) *scaling-down* when cluster density remains high and centrality decreases as terms are less and less adopted. Hence, we defined the topics in Quadrant ① as emerging niche clusters that could scale-out, becoming *promising topics* (clusters) when reaching Quadrant ②. If a niche cluster increases its internal structuration, it scales-up to the mainstream in Quadrant ③ characterised by motor topics. Quadrant ④ hosts scaling-down for fading research topics.

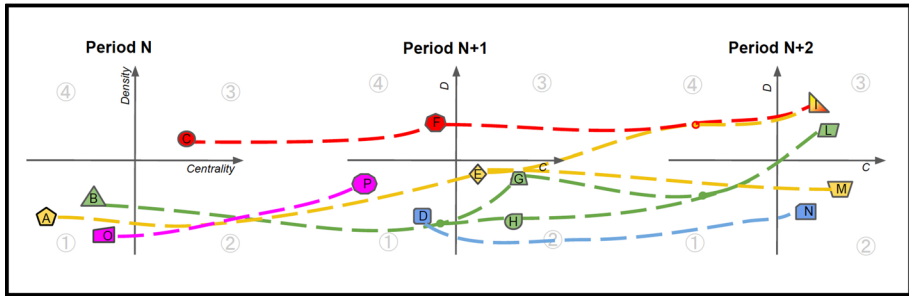
Enhanced with the concepts of *niche* and *mainstream*, this revisited SD provides a picture of the structure of a field as time passes, with special attention to the development of semantic clusters from niche to mainstream. But a single SD is not enough to reveal the underlying dynamics when clusters move from one quadrant to another over time. Especially for a large time-span, this overall picture could over-represent the time periods of more intense research production and under-represent the others. A temporal bias would therefore particularly favour recent decades characterised by an increase in the number of publications. As observed by Magrini et al. (2019) on a corpus of scholarly publications on legumes over 1980–2018, nearly 70% were published between 2000–2018. That is why we moved towards a diachronic approach for revealing the cumulative knowledge development over time that is a main characteristic of lock-in phenomenon.

### An original approach of the life cycle of semantic clusters: from a synchronic to a diachronic SD perspective

To move from a static view of the revisited SD (Fig. 2) to a temporal dynamic one, we propose to take the whole period under study and slice it into time spans ( $N, N+1, N+2$ , etc.), then conduct the co-word analysis separately in order to generate a SD by time span. These temporal SDs provide what we called a *synchronic* perspective of semantic clusters positions through time as illustrated in Figs. 3 and 4.

As topic clusters can change their internal composition over time (new terms appear, others disappear), it is impossible to reveal the evolution path of each topic just by looking at the *synchronic strategic diagram*. To move from a *synchronic* to a *diachronic* perspective, we proposed a final step by incorporating the contribution of Chavalarias and Cointet (2008, 2013) who introduced a method for intertemporal matching of topics over time called *phylogenies*. Cluster matching over different time periods is based on similarity calculation (on words composing clusters), making it possible to identify temporal series of





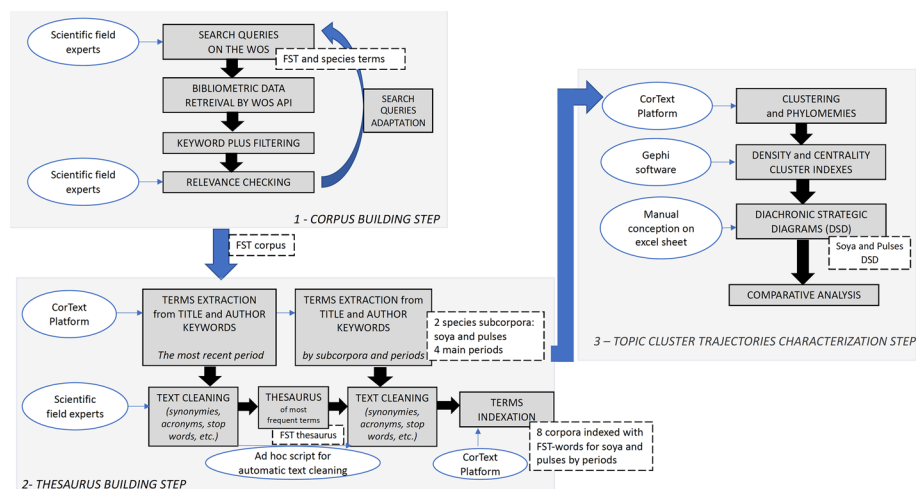
**Fig. 4** The Diachronic Strategic Diagram perspective over the period  $[N, N+2]$  (authors). the Diachronic Strategic Diagram reveals the links between the SDs mapped on different periods, thanks to intertemporal matching between clusters (represented by a capital letter and dotted arrows). One can observe various life-cycle paths (i.e., trajectories) of semantic clusters revealing strong or low accumulation over time

close semantic clusters. The resulting network is called *Phylomemetic network* in analogy with evolutive biology: inheritance patterns describe the dynamics of transformation in a scientific field through topic emergence or decline, merging, or splitting.<sup>3</sup>

Through phylomemetic links between SDs over several periods of time, one can depict the life cycle of a scientific field. One can infer different research paths through the position of clusters in the Strategic Diagram over time: scientific fields, able to rapidly develop niche clusters towards mainstream clusters, benefit from increasing returns of adoption mechanisms meaning stronger cumulative knowledge development. In contrast, scientific fields that do not achieve such dynamics receive less attention and remain in niche clusters, benefiting from less knowledge development. We hypothesise that such different dynamics over a long time period could reveal a lock-in situation in research activity. On the one hand, a research field with topics more frequently characterised as mainstream will sustain more knowledge development; this would then most likely foster more innovation. On the other hand, a research field with more niche topics will not gain enough interest; this will weaken the development of the knowledge base for that particular field, and probably undermine innovation processes. Such a research field benefiting from less cumulative effects over decades compared with a field from which theme clusters (topics) are more frequently characterised as mainstream, can be described as being locked-in.

We propose to apply this original approach to two fields competing over time, pulses and soya (belonging to the same botanic family: legumes) in order to assess if the lock-in situation that is observed in production and consumption concerning more cumulative effects on soya than on pulses is also observed through their scientific dynamics in the food and sciences technology (FST) field.

<sup>3</sup> As introduced by Chavalarias and Cointet (2013:1): “the concept of ‘phylomemetic network’ is used by analogy to biological phylogenetic trees, which account for evolutionary relationships between genes. We do not make any assumption concerning the type of dynamics underlying the evolution and diffusion of terms. As such, contrarily to previous works in line with the memetics theory, which have already coined the term, we do not claim that cultural entities (memes) evolve following the same laws of selection as biological replicators (genes) do”.



**Fig. 5** Main steps from retrieval to topic cluster trajectories analysis

## Materials and methods

This section introduces the data retrieval and corpus building steps, and the data processing employed to extract and normalise relevant terms for the analysis based on strategic diagrams. Figure 5 sums up the methodology we followed. Data processing and thesaurus construction were implemented using the CorText Manager platform, a free online platform for text analysis,<sup>4</sup> and an ad hoc Python script. Clustering and phylomemies were also generated using Cortext, while the software Gephi and Excel were used to analyse the results.

### Data retrieval and corpus building

Four search queries were designed to retrieve a set of publications (articles, books, book chapters, and reviews) from the Web of Science platform. With the help of experts in FST we defined a query for each of the following sub-themes of FST: Processing, Nutrition, Allergy, and Sensory. These queries were used in a study by Magrini et al. (2019) that retrieved a global corpus over legumes in various fields of research in order to compare the quantity of publications on soya and pulses over time. Note that these search queries excluded soya oil topics. Protein is definitely becoming a major driver of crop development compared to oil (EU, 2018). For the current study, we extended the publication dates already covered to 1957–2017 in order to get a better assessment of life-cycle dynamics.

The methodology we adopted to build the corpus on FST (based on the publication records retrieved from the WoS on pulses and soya) is available from Magrini et al. (2019). In a nutshell:

<sup>4</sup> <https://www.cortext.net>.

- (i) we applied an indexing process to remove any *Keywords Plus*<sup>5</sup> in order to retain bibliographic records only using the same relevant terms as in search queries amongst titles and authors' keywords. As a by-product, we indexed each record retrieved as soya or pulse-focused or both. We used titles and authors' keywords only and not abstracts, as abstracts were not indexed by the WoS before 1990. In addition, one expects that authors use the most important terms of their research in their titles and keywords.
- (ii) we checked and controlled the relevance of the records retrieved: 8 researchers on FST scrutinised the corpus using an in-house bibliometric platform to establish excluding conditions through WoS categories (WC) or journal delineation. This step was discontinued when less than 15% among a sample of 300 randomly selected records were deemed irrelevant. The check and control steps with experts in the field led to a representative corpus on FST, as opposed to some bibliometric studies conducted without either close collaboration between experts or *Keyword Plus* filtering.

The list of 39,036 record identifiers (UT codes) composing our FST corpus<sup>6</sup> observed from 1957 to 2017 is publicly available.<sup>7</sup>

## Data processing of terms in the FST corpus

We listed and standardised the most frequent terms from titles and authors' keywords<sup>8</sup> according to four time periods, in order to avoid biases related to the uneven distribution of publications through time and between species.

The initial main step consisted in splitting the corpus into four adjacent time periods characterised by a balanced publication number: 1956–1989, 1990–1999, 2000–2009, and 2010–2017, with the CorTexT '*Period Slicer*' script<sup>9</sup>; and extracting the 1,000 most frequent multi-terms (noun phrases composed of 1 to 3 terms) from titles and authors' keywords over the 2000–2017 period (the most productive period) to list relevant terms. This task was performed by using the '*Terms Extraction*' script of CorTexT, which performed tokenisation and stemming of multi-terms; and an approach similar to van Eck et al. (2010) to eliminate irrelevant terms. For instance, common terms such as "study", "analysis", etc. were removed. As in Rezaeian et al. (2017), text-mining was supplemented by expert evaluation to further filter irrelevant terms (non-informative or non-pertinent) and to improve the term standardisation run by CorTexT. For instance, experts helped in regrouping synonyms and terms related to the same subject or by matching developed terms with acronyms. This step resulted in, on the one hand, a thesaurus of relevant terms that we called "FST standardised terms", and on the other hand, a thesaurus of irrelevant terms (called

<sup>5</sup> Keywords Plus are keywords algorithmically added by the WoS and that differ from the author keywords.

<sup>6</sup> The distribution of FST sub-themes (Allergy, Nutrition, Processing, Sensory Analysis) shows that the most important subtheme concerns processing methods and food applications, indexed for 83% of the records (32,376 records). Then comes the Nutrition subtheme, with 12975 records, Sensory Analysis with 4416 mentions and finally Allergy with 1594.

<sup>7</sup> <https://doi.org/10.15454/JE7YY4>

<sup>8</sup> As abstracts were not available before 1991, we only considered terms from authors' keywords and titles as units of analysis.

<sup>9</sup> <https://docs.cortext.net/period-slicer/>.

also stop-words). We hence constructed a standardised thesaurus based on legumes in FST (soya and pulses).

The second main step consisted in splitting corpora between soya and pulses, using the Cortext ‘*Query*’ script (records mentioning both soya and pulses were affected to each subcorpus and represent only 1,419 records) to extract from each one the 1,000 more frequent multi-terms for the four time-periods considered above. Hence, eight lists of terms were obtained and merged into a single thesaurus, containing therefore 8,000 multi-terms (after lemmatisation processed by CorText and according to the thesaurus created in the first step). We then designed a specific algorithm to cleanse the extracted terms, according to the list we constructed at the first step (above). That is, we removed all stop-words, duplicate entries and matched synonyms. This lemmatisation overseen by experts reinforced the relevance of the terms we retained for the analysis. Finally, experts performed a visual inspection to check the final list of terms and enrich the stop-words, and adjust the thesaurus describing FST on legumes. After the experts’ validation, the final thesaurus of most frequent terms is composed of 936 terms (main forms), that we considered as being the main terms describing the research output on soya and pulses in the FST field over the period 1956–2017. These terms are called “FST-words” for the following analysis.

The third step consisted in indexing each record from our corpus with the FST-words with ‘*Terms Indexation*’ CorText script. Hence, the FST-words represent a new metadata associated with records, which we considered as our unit of analysis for the semantic clustering to be processed.

### FST co-words clustering and mapping through strategic diagram

We employed the ‘*Network Mapping*’ script from Cortext to map the networks of FST-words. This script maps and analyses networks of heterogeneous and homogenous nodes. To build the co-words network, we selected the FST-words as nodes. We chose to map the 250 most frequent nodes, separately for soya and pulses. Edges were obtained by co-occurrences of FST-words in each record, their proximity was measured by the *cosine* function. To eliminate trivial edges, we considered only those connecting the seven top neighbours of each node. This parameter was set after running various tests, and results as the best trade-off between noisy data reduction and information loss. To identify clusters of strongly connected words, we applied the ‘*Infomap*’ algorithm for community detection (Rosvall and Bergstrom, 2008). A comparison using the ‘*Louvain*’ algorithm showed that *Infomap* resulted in finer-grained clusters. Clusters representing research topics were automatically labeled by CorText with the two most connected words (*degree centrality*). Then, based on the clusters established by timespans, we tracked the evolution of main clusters thanks to intertemporal matching with phylomemories. The phylomemories between clusters over time are established by a ‘*Sankey diagram*’ on Cortext, recently used by Marvuglia et al. (2020) and similar to the ‘*river networks*’ used by (Rule et al., 2015).

Network maps were then exported to Gephi (Bastian et al., 2009) to compute cluster density and centrality indexes (Callon, 1991). Different centrality metrics exist; we chose the *closeness centrality* (Jackson, 2010) that measures the average of the shortest path lengths from one node to all other nodes in the network. Data were then extracted and, using Excel software, a Strategic Diagram was mapped for each timespan, for Soya and Pulses separately. As the origin in each SD, we took the mean of the values observed (both for centrality and density indexes) in each timespan. We plotted the clusters in each SD by time period in order to visualise the Synchronic Strategic Diagram (SSD) perspective.

**Table 1** Frequencies and growth rate of records indexed for pulses and soya by decade over 1956–2017 in the FST corpus (retrieved from the WoS)

Time span	1956–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2017	Full period
Soya	323	1033	1626	4563	7904	9776	25,225
	–	+ 220%	+ 57%	+ 181%	+ 73%	+ 24%	
Pulses	244	558	1277	3194	3946	5346	14,565
	–	+ 129%	+ 129%	+ 150%	+ 24%	+ 35%	

Hence we obtained two SSDs: one for soya and another for pulses. For the final step we identified the life cycle of clusters through the phylomemy in order to represent the Diachronic Strategic Diagram (DSD) perspective for soya and for pulses presented in the Results section.

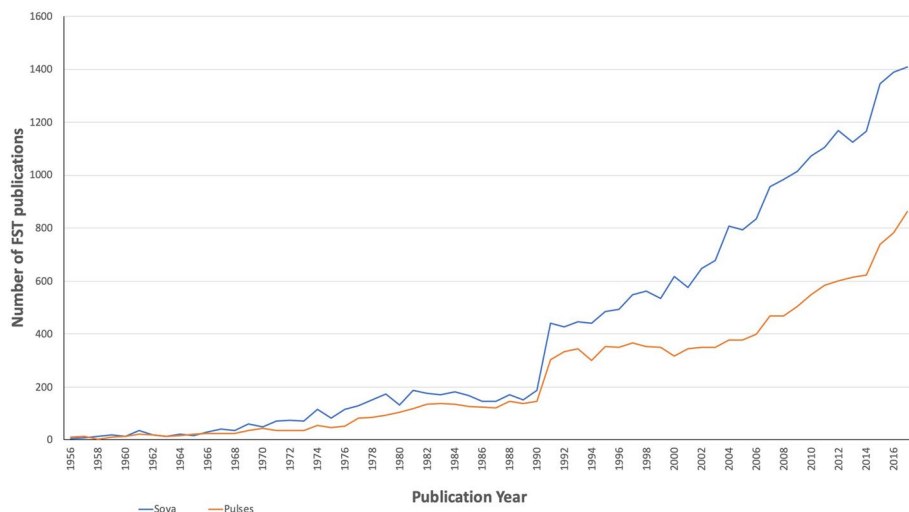
## Results and Discussion

We first described the corpus through the shares of species among records and their evolution over time. Secondly, we presented the findings of the co-word analysis on FST-words for soya and for pulse records at each of the 4 timespans. The intertemporal-matching of topics (phylomemies) reveals the different dynamics of the FST for soya and pulses over time. Next, the calculation of density and centrality indices allows to plot topics (semantic clusters) on synchronic strategic diagrams (SSD) for soya and for pulses at various timespans. As a result we are able to establish a diachronic perspective to characterise the topics' trajectories and assess, in general, a lock-in situation for pulses. We discussed the future work that these original results open both as regards to ways to assess lock-in in science and to develop a sound science policy to unlock pulses.

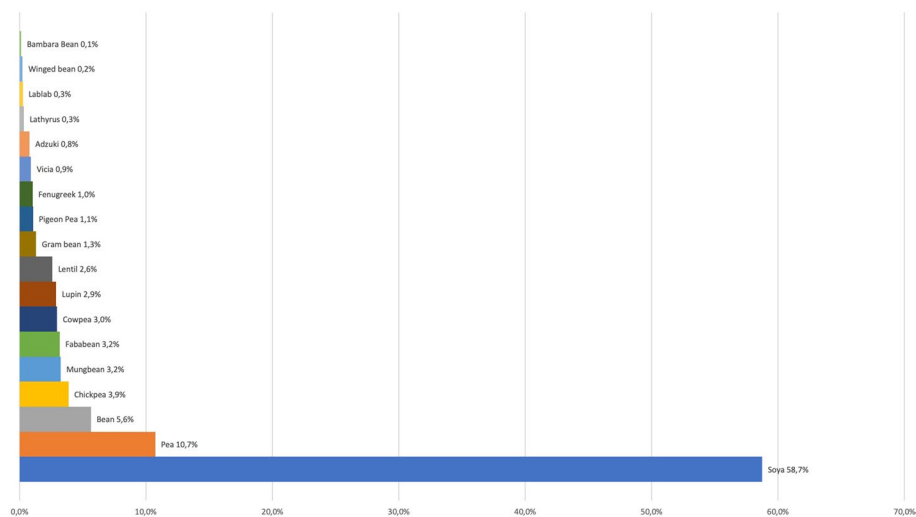
### Record frequencies over time: an imbalanced investment of soya and pulses in FST

Both species are concerned by the increase of publications, particularly from the 21st century, but the share of species (Table 1) suggests that research output is imbalanced in favour of Soya (25,225 records vs 14,565 on Pulses). Moreover, this imbalance is higher than in the general analysis of Magrini et al. (2019) over various fields. The two subcorpora species only overlap slightly, as only a few records were indexed with both soya and pulses (1,437 over the period). Soya represents the majority of the corpus throughout the considered time period. Whilst the gap between soya and pulses publications is not that high, with a strong increase of papers during the 1960s, this gap rose over the remaining period as the number of papers on soya was almost double those of pulses.<sup>10</sup> However, a slight inflection occurred during the last decade with a higher annual growth of publications for pulses (7%) compared with soya (4%), while the annual growth rate in the entire WoS collection is equivalent to 5–6% (Johnson et al. 2018). Figure 6 depicts this unbalanced evolution

<sup>10</sup> Note also that a large quantity of research publications is observed in the 1990s (as in the entire WoS Core collection) partly due to the inclusion, from 1990 on, of both abstracts and authors' keywords when the WoS indexed records, leading therefore to larger sets of retrieved records, as explained in Sect. 3.



**Fig. 6** Record frequencies curves for soya and pulses over 1956–2017



**Fig. 7** Mentions of legume species in the FST corpus over 1957–2017 (through titles and authors' keywords)

reinforced over the last decades, even though a slight inflection occurs in the speed of growth between soya and pulses in the last years.

When considering the various main pulse species (Fig. 7), the predominance of soya is even stronger: soya represents nearly 60% of records while the second legume species is pea with only under 11% of records. All other main pulses (even if well-known world-wide such as beans, chickpea or lentils) appear as minor species in FST research.

These statistics may partly explain the lock-in situation for pulses observed on markets (see the Introduction section) compared to soya which benefits from a stronger position

in the FST field (and as observed in other research fields accounted for in Magrini et al., 2019). The objective of our study was to depict this lock-in in a new way through the dynamics of the semantic clusters over time. Applying the method explained in the “[Materials and methods](#)” Section, we moved to a diachronic strategic diagram perspective of semantic clusters for soya and pulses which we shall comment hereafter.

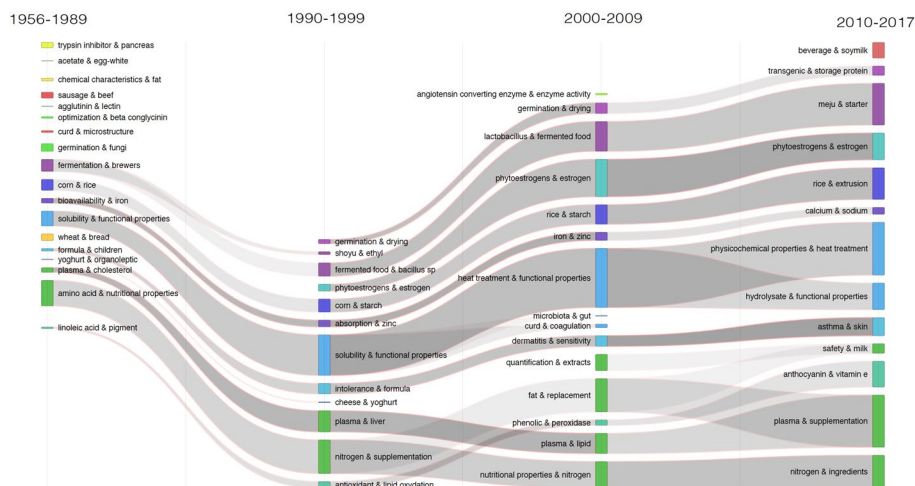
### The topic clusters evolution through phylomemories: soya vs. pulses

As explained in the “[Materials and methods](#)” Section, to map the synchronic strategic diagrams (SSD), respectively for soya and for pulses, and over the four time spans considered, we generated phylomemories visualised with ‘*Sankey Diagrams*’ (Figs. 9 and 10). The clusters are connected over time when many terms are shared between them; the nuances of dark shading show how strong the connection is. Tube width is proportional to the number of records related to the topic. Each cluster is automatically labeled by their two most connected terms. Before moving to the analysis in terms of strategic diagrams, the discussion of some features of phylomemories revealed interesting insights about the dynamics of knowledge in FST relating to soya and pulses.

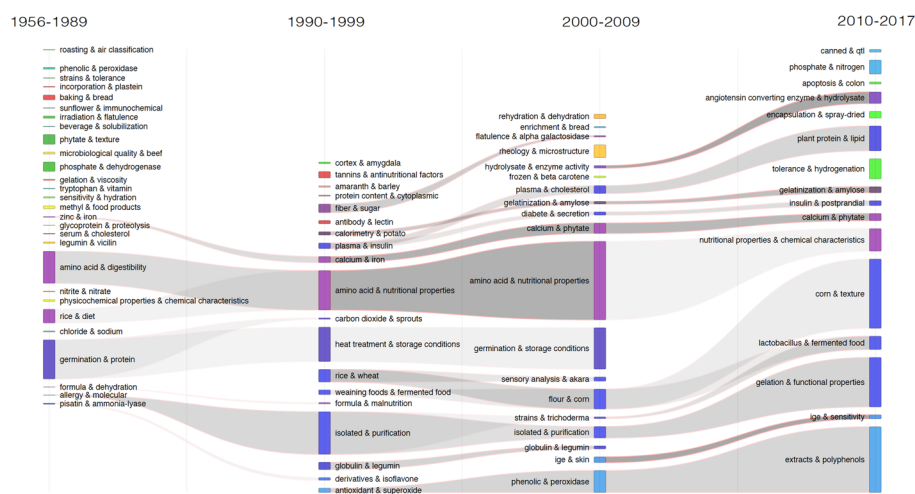
First, we observed that soya has a lower number of clusters than pulses by periods: around 14 to 18 clusters for Soya against 20 to 29 for pulses, which could mean a higher degree of scientific structure on soya compared to pulses. Pulses are characterised by a larger semantic dispersion, and by more isolated clusters—clusters which are not linked to other clusters in the preceding or following time period. It means that the set of terms of isolated clusters is very dissimilar compared to those of clusters at different time periods. Nevertheless, whilst some isolated topics do not survive through time, some terms that constituted them could reappear in new clusters in another period. However, when comparing soya and pulses we observe that during the first period, half of the clusters were isolated and counted few publications both for soya and pulses, while in the other periods only two clusters were isolated for soya and much more for pulses (at least 4 per period). Thus, this first description of semantic clusters over time indicates that soya seems to benefit from more structured topics than pulses do.

Secondly, we observed that almost all soya topics from the first period (1956–1989) which were maintained in the next period (1990–1990), survived throughout the remaining period: eight of them were at the origin of long topics’ series. This shows that the foundations of the FST on soya were already well established since the first considered period. We also observe that after the 1990s the research field on soya underwent very few transformations: we observed a degree of robustness of the connected clusters, and only limited events of *declining*, *branching* or *emerging* topics. This confirms that FST research on soya appears to follow a dynamics, characterised by an accumulation of records through close semantic clusters. Conversely, the structure of the research field of pulses appears less structured as we observe a few long temporal series of topics: most of them started during the 1990s and only three of them originated from the first period. Compared to soya, the field of pulses underwent many more transformations. In other words, pulses phylomemories seem less robust through time, meaning that their evolution is characterised by a stronger mutation in the set of terms at successive time periods. Looking at the number of related records, some topics for pulses presented a strong fluctuation during their evolution and not a simple incremental path; some topics declined or emerged during the second or the third period, and we also observe more events of *branching* or *recombination* of topics through





**Fig. 8** Soya topics evolution visualised by Sankey Diagrams over four timespans (1956–1989; 1990–1999; 2000–2009; 2010–2017) (CorText platform)



**Fig. 9** Pulses topics evolution visualised by Sankey Diagram over four time spans (1956–1989; 1990–1999; 2000–2009; 2010–2017) (CorText platform)

time. These results indicate a temporal delay in the construction of the FST on pulses compared to soya, as topics seem to follow less stable and less cumulative trajectories.

Having described the general evolution of semantic clusters over time for soya and pulses through phylomemories, we moved to a deeper perspective by characterising the trajectory of those clusters according to their position in strategic diagrams revisited through niche and mainstream concepts (see Fig. 2): we called this original perspective



“the diachronic strategic diagram” (DSS) formalised in the “[Analysing scientific knowledge dynamics through text mining: a renewed approach](#)” Section.

### The diachronic strategic diagram perspective: characterising the trajectories of topic clusters of soya vs pulses

The intertemporal-matching of clusters, visualised as phylomemies (Figs. 8 and 9), underlies links between several clusters at different timespans. It enables us to track temporal series of topic clusters, which could be considered as research-paths characterised by accumulative records and subtopics (according to degrees of density and centrality). These temporal series demand a deeper analysis, to reinforce our analysis of a lock-in trend (i.e., a less accumulated knowledge path) that could characterise pulses. Indeed, whilst the results discussed above show a less structured research path for pulses, as we observe numerous isolated clusters at each period, analysing temporal series could reveal more cumulative research paths for the clusters constituting them. Therefore, the analysis hereafter focuses only on clusters linked in a temporal series (that is, linking at least two timespans).

To conduct this analysis we projected the clusters into strategic diagrams by timespan (what we call Synchronic Strategic Diagram perspective in “[Analysing scientific knowledge dynamics through text mining: a renewed approach](#)” Section) to obtain Figs. 10 and 11. In those figures, the temporal series of topics are presented in rows; for each cluster we inserted a pictogram representing its position into the quadrants of the SSD (according to its indexes of density and closeness centrality, which are reported too); we reported labels identifying clusters in phylomemies and the number of records related to each of them; we grouped the temporal series of clusters by colour-naming them according to a capital letter. The last column sums-up the trajectory of clusters according to their successive position in the SSD, leading to what we called the diachronic strategic diagram perspective.





First of all, when considering clusters which are not isolated, that is to say, following a trajectory that could be interpreted in terms of life cycle over the period, we counted 9 series for soya and 12 for pulses. Hence, the research dispersion still appears higher for pulses than for soya.

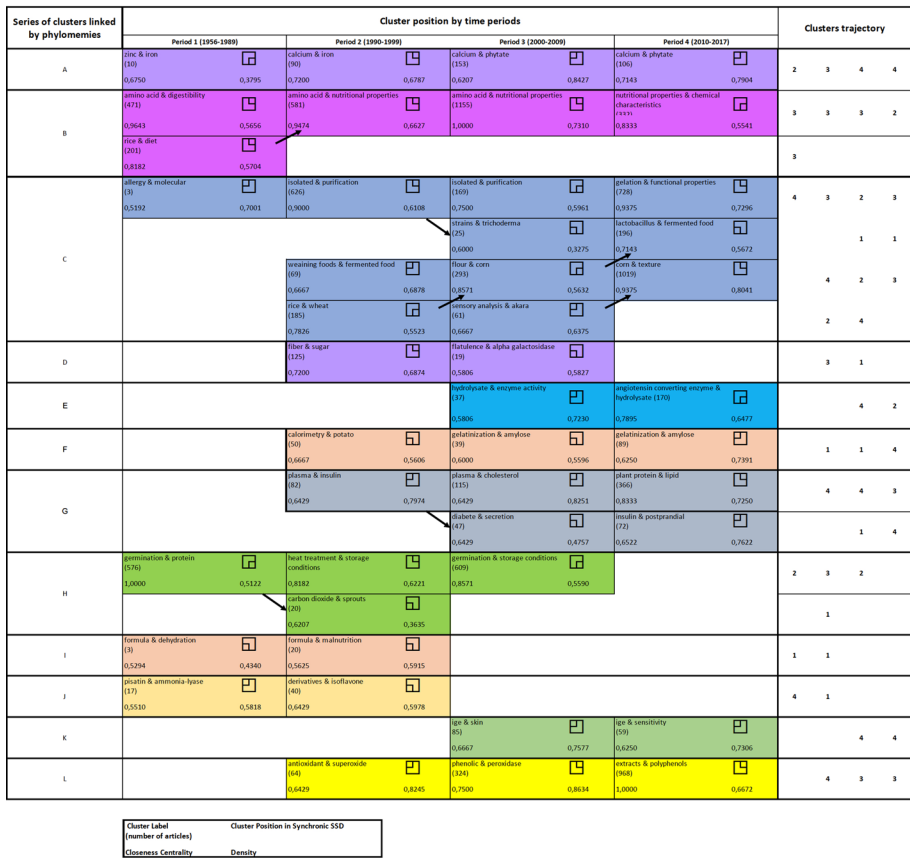
Secondly, when looking at the characterisation of these clusters according to their position in the strategic diagram (Table 2), we observed a higher number of mainstream clusters for soya than for pulses. Particularly, as regards the number of clusters positioned as motor ones. Pulses also present more emerging clusters and, more particularly, when comparing trajectories of pulse and soya clusters (Figs. 10 and 11) we observe that all the emerging clusters remained so or disappeared<sup>11</sup> in successive timespans, whilst this was never the case for soya. This point is most important as it reveals an incapacity to reach sufficient cumulative research outputs (here observed as scholarly publications), a critical mass that could allow higher knowledge development (favouring emerging or branching clusters) and reinforcement of previous knowledge (characterising, in our sense, motor and then fading clusters). Indeed, we also observed that soya cluster trajectories presented 8 temporal series characterised by clusters maintaining their *motor* or *fading* positions, or moving toward, respectively, *fading* or *motor* positions, through at least three successive timespans, whilst this was the case for only 4 series with pulses (series A, B, G and L in

<sup>11</sup> In one case an emerging cluster moved directly to the fading position for pulses (temporal series F in Fig. 12).

Groups of clusters linked by phylomemes	Cluster position by time periods				Clusters trajectory			
	Period 1 (1956-1989)	Period 2 (1990-1999)	Period 3 (2000-2009)	Period 4 (2010-2017)				
A	fermentation & brewers (316) 0.8947	fermented food & bacillus sp (348) 0.7333	lactobacillus & fermented food (775) 0.7778	meju & starter (1084) 0.8	3	4	3	4
		shoyo & ethyl (75) 0.785714				3		
B	corn & rice (285) 0.85	corn & starch (340) 0.9167	rice & starch (508) 0.7	rice & extrusion (823) 0.8571	2	2	4	3
C	bioavailability & iron (125) 0.7727	absorption & zinc (176) 0.6345	iron & zinc (214) 0.7283	calcium & sodium (190) 0.8	3	4	4	2
D	solubility & functional properties (387) 0.85	solubility & functional properties (1054) 0.6128	heat treatment & functional properties (1528) 0.8187	physicochemical properties & heat treatment (1365) 0.9231	3	3	3	3
				hydrolysate & functional properties (690) 0.9231				2
			microbiota & gut (19) 0.6087					1
			card & coagulation (92) 0.5833					1
E	formula & children (65) 0.6071	intolerance & formula (275) 0.7473	dermatitis & sensitivity (279) 0.895352	asthma & skin (471) 0.9108	4	4	4	4
F	plasma & cholesterol (123) 0.7727	plasma & liver (351) 0.7915	plasma & lipid (527) 0.83338	plasma & supplementation (1356) 0.8189	3	3	3	3
			fat & replacement (866) 0.875			2		
	amino acid & nutritional properties (666) 0.9444	nitrogen & supplementation (877) 0.5712	nutritional properties & nitrogen (692) 0.7606	nitrogen & ingredients (856) 0.662	3	3	3	4
				safety & milk (250) 0.8571				2
G	linoleic acid & pigment (45) 0.6296	antioxidant & lipid oxidation (172) 0.5796	phenolic & peroxidase (141) 0.6364	anthocyanin & vitamin e (671) 0.8571	4	4	4	2
H		germination & drying (112) 0.7333	germination & drying (227) 0.8235	transgenic & storage protein (243) 0.7059	1	2	1	
I				coverage & soybean (406) 0.7209				3
J		phytoestrogens & estrogen (179) 0.6111	phytoestrogens & estrogen (573) 0.8298	phytoestrogens & estrogen (606) 0.7778	4	3	4	

Cluster Label (number of articles)	Cluster Position in Synchronic SSD
Closeness Centrality	Density

**Fig. 10** Diachronic Strategic Diagram Perspective: Soya clusters trajectories (1957–2017) (authors). Rows correspond to series of intertemporal linked topics over four time spans. The names and colours of the series correspond to those visualised in Fig. 8. The first column indicates a capital letter to identify each series. The following characteristics describe each cluster: its name (corresponding to the two terms having the higher degree of centrality) and number of records; density and closeness centrality indexes that led to position the cluster in the Synchronic Strategic Diagram; a pictogram symbolises this position:  emerging cluster (position 1),  promising cluster (position 2),  motor cluster (position 3),  fading cluster (position 4). The diachronic strategic diagram (DSD) perspective column sums up those successive positions held by clusters, that is the topics' trajectory over time represented by the sequence of quadrants' number in the Strategic Diagram at each period



**Fig. 11** Diachronic Strategic Diagram perspective: Pulses clusters' trajectories (1957–2017) (authors). see Fig. 10

**Table 2** Position of clusters belonging to a temporal series (from Figs. 11 and 12) according to the strategic diagram for soya and for pulses

	1—Emerging	2—Promising	3—Motor	4—Fading	Total
Pulse cluster frequencies (%)	10 (22%)	8 (18%)	13 (29%)	14 (31%)	45
	<i>Niche position</i>		<i>Mainstream position</i>		
	18 (40%)		27 (60%)		
Soya Cluster frequencies (%)	4 (9%)	9 (20%)	17 (38%)	15 (33%)	45
	<i>Niche position</i>		<i>Mainstream position</i>		
	13 (29%)		33 (71%)		

Fig. 11). In short, soya clusters presented a stronger accumulative research path by cumulating *motor* and *fading* positions (i.e., mainstream positions) over time, while research on pulses did not meet with the same success.

In other words, whilst for both species many promising topics were detected, we observe that their capacity to scale-up and enter the mainstream position was higher for soya.

## Conclusion and research directions

Research activities are essential for sustainability transition, and lock-in in science per se could prevent sustainability transition if scientific knowledge is less developed in areas considered as essential. Nevertheless, analysing lock-in in science is challenging. We developed an original approach, as a first attempt, that we applied to legumes literature in FST, from which we questioned how we can assess or not a lock-in situation in science, opening a new research agenda to pursue this analysis.

Firstly, as lock-in deals with long-term path-dependency, we revisited the strategic diagram (SD) of Callon with a diachronic perspective revealing the trajectories of semantic (i.e., topic) clusters from *niche* to *mainstream* positions, measured through centrality and density indices. By combining phylogenies and the SDs we were able to depict the main trajectories of research topics for assessing lock-in, viewed as an incapacity to reach mainstream levels of clustering. We were thereby able to reveal both the internal evolution of topics and their position within the field according to their capacity to reach, or not, mainstream positions. The analysis performed showed that characterising clusters separately from each other is limited. Even if cluster positions in SD by periods of time (SSD) reveal insights as regards to strong research investment (when numerous clusters are in mainstream positions, ie. the quadrants 3 and 4 of the SD) compared to another field characterised by less clusters, analysing trajectories remains essential. The ability of clusters to survive over time and be positioned in mainstream positions reveal long-term accumulation in those topic clusters and therefore, higher research consolidation; as well as a low number in isolated clusters (i.e., clusters which are not linked to other clusters over time and therefore, without a consolidated trajectory). Isolated clusters reveal that research is also an exploration of certain topics that might, after initial investigation, be considered as non-fruitful compared with other topics. Isolated clusters are expected as part of normal science, but when these isolated clusters grow in numbers over time, this could also reveal an incapacity to stabilise research activities in developing topics.

Secondly, this work needs to undergo further tests, both by applying it to other research fields and varying the parameters. Indeed, the assessment we drew on lock-in for pulses compared to soya remains dependent on the indicators we chose to position clusters in the SD. Introducing variations in the thresholds used for measuring centrality and closeness indexes could generate complementary information. In addition, variations in time segmentation could also be considered. The proposed workflow (algorithms run and figures crafted) is considered for inclusion in the Cortext text-mining platform in order to enhance reproducibility and foster its application to other research fields.

Thirdly, whilst co-word analysis remains the core analysis to analyse the evolution of scientific topics, analysing the authors and/or institutions networks should allow to take this further, and perhaps to provide an explanation in the path-dependency process. For instance, niche trajectory persistence over time could be linked to low density and/or centrality researcher networks. Analysing both social and semantic networks (e.g., Roth and

Cointet, 2010) would provide additional insights. Questioning also the funding of researchers on soya and pulses could reveal different amounts of investments through topics, particularly between public and private funds.

Fourthly, as regards to science policies, these tools could help policy-makers better appreciate science trajectories and detect areas of research that need to be reinforced compared to certain topics that are over-studied or that benefited from a more accumulative dynamic. In particular, our analysis suggests that pulses encounter difficulties in stabilising a research path that could gain more centrality and density, as topic clusters on pulses are less connected. This could also be linked to the fact that the botanic family of pulses has not yet reached a level of common interest to structure common topics in which increasing returns adoption mechanisms could engage (e.g., the more knowledge is developed on certain topics, the more research is devoted to these topics). Hence, this brings up the question whether the diversity of pulses species creates specific difficulties in structuring increasing returns on pulses research, and therefore which science policy must be developed to favour more biodiversity in research.

But concerning specifically the research field on which we applied the method developed, further questions remain given that a strong driver of soya development is oil outlet (being for food, fuel or feed). Even if we excluded the oil theme from our bibliographic corpus (for comparison between soya and pulses) part of the cumulative output on soya could be explained by also a cumulative research output on soya oil theme that we did not measure. Soya is characterised by a multifaceted business model (the so-called “soy complex”) that could notably explain the availability of more fundings for this research. Therefore, further exploration must be considered to also unpack the links between major fields of research, and links between *markets* and *research*. In other words, this question remains challenging: *does lock-in within research is the consequence of lock-in within markets?*

Besides, apart from that question, another major one concerns the “locking cracking strategy”. The analysis of the sources of research fundings can be a first step in order to assess how public research could investigate fields of research currently locked-in. The main idea is that public research could reinforce topic emergence providing new insights which pave the way for new directions, particularly for societal transformation. When considering other fields such as, for instance, transport technologies, do we observe a cumulative path for topics linked to fossil-based technologies over decades while the ones on alternative technologies remained locked-in until a certain period, given the fact that nowadays electric vehicles are more developed on markets? Do we observe changes in fundings regarding those two main fields of research? Do specific fundings in field research locked-in could change the path-dependency process, and contribute to overcome the lock-in observed on markets?

Other research fields could be considered through this competing technologies framework linked to lock-in issue, to develop measurements of research trajectories and strategies to deviate those trajectories; considering that such an analysis must be conducted over large period of time given that cognitive dissonance mechanism takes time in cracking lock-in.<sup>12</sup> In addition to funding, the choice of subjects and research fields to be funded is also challenging. Again, we refer to the literature from *Transition Studies* suggesting that

<sup>12</sup> New paths can be hampered by the mechanism of “cognitive dissonance (...) between the potential of the new and the security of the old”, that is “the greater the distance between a novel solution and the accepted one, the larger is the lock-in to previous tradition. And so a hysteresis—a delayed response to change—exists” (Arthur, 2009:139–140).

front-running topics could emerge from experiments by niche-actors which are not well established networks but that constitute first seeds on which research can investigate to go further. In that way, cracking lock-in questions also the ways new topics for research emerge and fit societal expectations. Analysing the topic trajectories throughout citizen networks and research ones could also constitute a way to investigate how lock-in could be overcome by societal pressure, and inversely.

**Acknowledgements** The authors acknowledge Hugues Leiser and experts from Food Sciences and Technology who helped build search queries on the WoS and the thesaurus dictionary: Marie-Jo Amiot-Carlin, Marc Anton, Jean-Michel Chardigny, Valérie Micard, Christophe Nguyen-Thé, and Stéphane Walrand. We kindly thank the CorText team; Tristan Salord for his help on certain scripts; and Alice Thomson-Thibault for her helpful comments and English editing of the manuscript. We also thank an anonymous reviewer whose comments and suggestions helped sharpen the argument.

**Author contributions** Conceptualisation, Data curation, Formal analysis: all authors. Funding acquisition: M-BM. Investigation: ML and M-BM. Methodology: all authors. Project administration, Resources: M-BM. Software: ML, GC. Supervision: M-BM and GC. Validation, Visualisation, Roles/Writing—original draft and Writing—review & editing: all authors.

**Funding** This work was supported by funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 727672 LEGVALUE (Fostering sustainable legume-based farming systems and agri-feed and food chains in the EU); from the Agence Nationale de la Recherche (ANR) under grant number ANR-11-LABX-0066; and the Occitanie region in France.

**Data availability** <https://doi.org/10.15454/JE7YY4>.

## Declarations

**Conflict of interest** Guillaume Cabanac serves in the Editorial Board member of *Scientometrics*.

## References

- Abdullah, M., Marinangeli, C., Jones, P., & Carlberg, J. (2017). Canadian potential healthcare and societal cost savings from consumption of pulses: A cost-of-illness analysis. *Nutrients*, 9, 793. <https://doi.org/10.3390/nu9070793>
- Adams, J., Hillier-Brown, F. C., Moore, H. J., Lake, A. A., Araujo-Soares, V., White, M., & Summerbell, C. (2016). Searching and synthesising ‘grey literature’ and ‘grey information’ in public health: Critical reflections on three case studies. *Systematic Reviews*, 5(1), 1–11.
- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of michigan Press.
- Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. Penguin Books.
- Bailón-Moreno, R., Jurado-Alameda, E., & Ruiz-Baños, R. (2006). The scientific network of surfactants: Structural analysis. *Journal of the American Society for Information Science and Technology*, 57, 949–960. <https://doi.org/10.1002/asi.20362>
- Balconi, M., Breschi, S., & Lissoni, F. (2004). Networks of inventors and the role of academia: An exploration of Italian patent data. *Research Policy*, 33, 127–145. [https://doi.org/10.1016/S0048-7333\(03\)00108-2](https://doi.org/10.1016/S0048-7333(03)00108-2)
- Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi : An Open Source Software for Exploring and Manipulating Networks. International AAAI Conference on Weblogs and Social Media 2.
- Borsi, B., & Schubert, A. (2011). Agrifood research in Europe: A global perspective. *Scientometrics*, 86, 133–154. <https://doi.org/10.1007/s11192-010-0235-3>
- Cahlik, T. (2000). Comparison of the maps of science. *Scientometrics*, 49(3), 373–387.
- Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22, 191–235. <https://doi.org/10.1177/053901883022002003>

- Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22, 155–205. <https://doi.org/10.1007/BF02019280>
- Callon, M., Rip, A., & Law, J. (1986). *Mapping the dynamics of science and technology: sociology of science in the real World*. Palgrave Macmillan Springer.
- Chavalarias, D., & Cointet, J.P., 2009. The reconstruction of science phylogeny. arXiv preprint <http://arXiv.org/0904.3154>
- Chavalarias, D., & Cointet, J.-P. (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study. *Scientometrics*, 75, 37–50. <https://doi.org/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, e54847. <https://doi.org/10.1371/journal.pone.0054847>
- Ciarli, T., & Ràfols, I. (2019). The relation between research priorities and societal demands: The case of rice. *Research Policy, New Frontiers in Science, Technology and Innovation Research from SPRU's 50th Anniversary Conference*, 48, 949–967. <https://doi.org/10.1016/j.respol.2018.10.027>
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, 5, 146–166. <https://doi.org/10.1016/j.joi.2010.10.002>
- Conti, C., Zanello, G., & Hall, A. (2021). Why are agri-food systems resistant to new directions of change? A systematic review. *Global Food Security*. <https://doi.org/10.1016/j.gfs.2021.100576>
- Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49, 1206–1223. [https://doi.org/10.1002/\(SICI\)1097-4571\(1998\)49:13%3c1206::AID-ASI7%3e3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-4571(1998)49:13%3c1206::AID-ASI7%3e3.0.CO;2-F)
- Courtial, J., Callon, M., & Sigogneau, A. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics*, 26, 231–242. <https://doi.org/10.1007/BF02016216>
- Cusworth, G., Garnett, T., & Lorimer, J. (2021). Legume dreams: The contested futures of sustainable plant-based food systems in Europe. *Global Environmental Change*, 69, 102321. <https://doi.org/10.1016/j.gloenvcha.2021.102321>
- Daim, T. U., Rueda, G., Martin, H., & Gerdts, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, Tech Mining: Exploiting Science and Technology Information Resources*, 73, 981–1012. <https://doi.org/10.1016/j.techfore.2006.04.004>
- Dosi, G., & Nelson, R. R. (2010). Technical change and industrial dynamics as evolutionary processes. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (pp. 51–127). Elsevier. [https://doi.org/10.1016/S0169-7218\(10\)01003-8](https://doi.org/10.1016/S0169-7218(10)01003-8)
- Drieger, P. (2013). Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*, 79, 4–17. <https://doi.org/10.1016/j.sbspro.2013.05.053>
- Epicoco, M., Oltra, V., & Saint Jean, M. (2014). Knowledge dynamics and sources of eco-innovation: Mapping the green chemistry community. *Technological Forecasting and Social Change*, 81, 388–402. <https://doi.org/10.1016/j.techfore.2013.03.006>
- European Commission. (2018). Report from the commission to the council and the European parliament on the development of plant proteins in the European Union, COM/2018/757. <https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:52018DC0757>
- Foyer, C. H., Siddique, K. H. M., Tai, A. P. K., Anders, S., Fodor, N., Wong, F.-L., Ludidi, N., et al. (2018). Modelling predicts that soybean is poised to dominate crop production across Africa: Soybean production in Africa. *Plant, Cell & Environment*, 42(9), 373–385. <https://doi.org/10.1111/pce.13466>
- Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy*, 33, 897–920. <https://doi.org/10.1016/j.respol.2004.01.015>
- Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (2019). *Springer handbook of science and technology indicators*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-02511-3>
- Guéguen, J., Walrand, S., & Bourgeois, O. (2016). Les protéines végétales : Contexte et potentiels en alimentation humaine. *Cahiers De Nutrition Et De Diététique*, 51, 177–185. <https://doi.org/10.1016/j.cnd.2016.02.001>
- Hallström, E., Carlsson-Kanyama, A., & Börjesson, P. (2015). Environmental impact of dietary change: A systematic review. *Journal of Cleaner Production*, 91, 1–11. <https://doi.org/10.1016/j.jclepro.2014.12.008>
- Havemeier, S., Erickson, J., & Slavin, J. (2017). Dietary guidance for pulses: The challenge and opportunity to be part of both the vegetable and protein food groups: Dietary guidance for pulses. *Annals of the New York Academy of Sciences*, 1392, 58–66. <https://doi.org/10.1111/nyas.13308>



- Heimeriks, G., & Boschma, R. (2014). The path- and place-dependent nature of scientific knowledge production in biotech 1986–2008. *Journal of Economic Geography*, 14, 339–364. <https://doi.org/10.1093/jeg/lbs052>
- Hotchkiss, J., & Potter, N. (1998). *Food science* (5th ed.). Springer.
- Hu, X., & Rousseau, R. (2018). A new approach to explore the knowledge transition path in the evolution of science & technology: From the biology of restriction enzymes to their application in biotechnology. *Journal of Informetrics*, 12(3), 842–857. <https://doi.org/10.1016/j.joi.2018.07.004>
- Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.
- Jallinoja, P., Niva, M., & Latvala, T. (2016). Future of sustainable eating? Examining the potential for expanding bean eating in a meat-eating culture. *Futures, SI: Futures for Food*, 83, 4–14. <https://doi.org/10.1016/j.futures.2016.03.006>
- Johnson, R., Watkinson, A., & Mabe, M. (2018). *The STM report. An overview of scientific and scholarly publishing* (5th ed.).
- Kuhn, T. S. (1970). The structure of scientific revolutions. In O. Neurath & T. S. Kuhn (Eds.), *International encyclopedia of unified science Foundations of the unity of science* (2nd ed., Vol. 2). University of Chicago Press.
- Lascialfari, M., Magrini, M.-B., & Triboulet, P. (2019). The drivers of product innovations in pulse-based foods: Insights from case studies in France, Italy and USA. *Journal of Innovation Economics*, 28, 111. <https://doi.org/10.3917/jie.028.0111>
- Lee, P.-C., & Su, H.-N. (2011). Quantitative mapping of scientific research—The case of electrical conducting polymer nanocomposite. *Technological Forecasting and Social Change*, 78, 132–151. <https://doi.org/10.1016/j.techfore.2010.06.002>
- Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. *Journal of Informetrics*, 5(3), 469–475. <https://doi.org/10.1016/j.joi.2011.01.008>
- Magrini, M.-B., Anton, M., Cholez, C., Corre-Hellou, G., Duc, G., Jeuffroy, M.-H., Meynard, J.-M., Pelzer, E., Voisin, A.-S., & Walrand, S. (2016). Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agri-food system. *Ecological Economics*, 126, 152–162. <https://doi.org/10.1016/j.ecolecon.2016.03.024>
- Magrini, M.-B., Cabanac, G., Lascialfari, M., Plumecocq, G., Amiot, M.-J., Anton, M., Arvisenet, G., Baranger, A., Bedoussac, L., Chardigny, J.-M., Duc, G., Jeuffroy, M.-H., Journet, E.-P., Juin, H., Larré, C., Leiser, H., Micard, V., Millot, D., Pilet-Nayel, M.-L., ... Wery, J. (2019). Peer-reviewed literature on grain legume species in the WoS (1980–2018): A comparative analysis of soybean and pulses. *Sustainability*, 11, 6833. <https://doi.org/10.3390/su11236833>
- Magrini, M.-B., Anton, M., Chardigny, J. M., Duc, G., Duru, M., Jeuffroy, M. H., Meynard, J. M., Micard, V., & Walrand, S. (2018). Pulses for sustainability: breaking agriculture and food sectors out of lock-in. *Frontiers in Sustainable Food Systems*. <https://doi.org/10.3389/fsufs.2018.00064>
- Magrini, M.-B., Salord, T., & Cabanac, G. (2022). The unbalanced development among legume species regarding sustainable and healthy agrifood systems in North-America and Europe focus on food product innovations. *Food Security*. <https://doi.org/10.1007/s12571-022-01294-9>
- Manners, R., & van Etten, J. (2018). Are agricultural researchers working on the right crops to enable food and nutrition security under future climates? *Global Environmental Change*, 53, 182–194. <https://doi.org/10.1016/j.gloenvcha.2018.09.010>
- Marvuglia, A., Havinga, L., Heidrich, O., Fonseca, J., Gaitani, N., & Reckien, D. (2020). Advances and challenges in assessing urban sustainability: An advanced bibliometric review. *Renewable and Sustainable Energy Reviews*, 124, 109788. <https://doi.org/10.1016/j.rser.2020.109788>
- Moed, H. F., Glänzel, W., Schmoch, U., Ziedonis, A. A., Valente, A., & Bassecoulard, E. (2004). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*. Kluwer Academic publishers.
- Peacock, M. S. (2009). Path dependence in the production of scientific knowledge. *Social Epistemology*, 23, 105–124. <https://doi.org/10.1080/02691720902962813>
- Peoples, M. B., Hauggaard-Nielsen, H., Huguenin-Elie, O., Jensen, E. S., Justes, E., Williams, M., et al. (2019). The contributions of legumes to reducing the environmental risk of agricultural production. In G. Lemaire (Ed.), *Agroecosystem diversity* (pp. 123–143). Elsevier.
- Pinto, A., Guerra, M., Carbas, B., Pathania, S., Castanho, A., & Brites, C. (2016). Challenges and opportunities for foodprocessing to promote consumption of pulses. *Revista de Ciências Agrárias*, 39(4), 571–582. <https://doi.org/10.19084/RCA16117>
- Poux, X., & Aubert, P.-M. (2018). An agroecological Europe in 2050: multifunctional agriculture for healthy eating. *Findings from the Ten Years for Agroecology (TYFA) Modelling Exercise Iddri-ASCA, Study*, 74, 9–18.



- Prabhakaran, T., Lathabai, H. H., George, S., & Changat, M. (2018). Towards prediction of paradigm shifts from scientific literature. *Scientometrics*, 117, 1611–1644. <https://doi.org/10.1007/s11192-018-2931-3>
- Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics*, 14(3), 101047. <https://doi.org/10.1016/j.joi.2020.101047>
- Rafols, I., Hopkins, M. M., Hoekman, J., Siepel, J., O'Hare, A., Perianes-Rodríguez, A., & Nightingale, P. (2014). Big Pharma, little science? *Technological Forecasting and Social Change*, 81, 22–38. <https://doi.org/10.1016/j.techfore.2012.06.007>
- Rezaeian, M., Montazeri, H., & Loonen, R. C. G. M. (2017). Science foresight using life-cycle analysis, text mining and clustering: A case study on natural ventilation. *Technological Forecasting and Social Change*, 118, 270–280. <https://doi.org/10.1016/j.techfore.2017.02.027>
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4), 1118–1123. <https://doi.org/10.1073/pnas.0706851105>
- 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, 1790–2014. *PNAS*, 112, 10837–10844. <https://doi.org/10.1073/pnas.1512221112>
- Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K., & Bloem, M. W. (2021). Legumes as a sustainable source of protein in human diets. *Global Food Security*, 28, 100520. <https://doi.org/10.1016/j.gfs.2021.100520>
- Sonnino, A. (2016). Leguminose da Granella e Ricerca Agricola - Pulses and Agricultural Research. Atti del Seminario Leguminose da Granella – Sant'Angelo Lodigiano, pp. 45–50. Retrieved Oct 14, 2016, from [https://sites.google.com/site/storiagricoltura/download-area/atti\\_seminari\\_mulsa](https://sites.google.com/site/storiagricoltura/download-area/atti_seminari_mulsa)
- Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. *Research Policy*, 33, 1615–1634. <https://doi.org/10.1016/j.respol.2004.09.008>
- Stegmann, J., & Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. *Scientometrics*, 56, 111–135. <https://doi.org/10.1023/A:1021954808804>
- van Eck, N., Waltman, L., Noyons, E., & Buter, R. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3), 581–596. <https://doi.org/10.1007/s11192-010-0173-0>
- Wasserman, S., & Faust, K. (1994). *Social network analysis methods and applications* (Vol. 8). Cambridge: Cambridge University Press. Retrieved from <https://doi.org/10.1017/CBO9780511815478>
- Weindl, I., Ost, M., Wiedmer, P., Schreiner, M., Neugart, S., Klopsch, R., Kühnhold, H., Kloas, W., Henkel, I. M., Schlüter, O., Bußler, S., Bellingrath-Kimura, S. D., Ma, H., Grune, T., Rolinski, S., & Klaus, S. (2020). Sustainable food protein supply reconciling human and ecosystem health: A Leibniz position. *Global Food Security*, 25, 100367. <https://doi.org/10.1016/j.gfs.2020.100367>
- Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L. J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J. A., De Vries, W., Majele Sibanda, L., ... Murray, C. J. L. (2019). Food in the Anthropocene: The EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet*, 393, 447–492. [https://doi.org/10.1016/S0140-6736\(18\)31788-4](https://doi.org/10.1016/S0140-6736(18)31788-4)
- Xu, H., Winnink, J., Yue, Z., Liu, Z., & Yuan, G. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics*, 14(2), 101014. <https://doi.org/10.1016/j.joi.2020.101014>
- Yang, Y., Wu, M., & Cui, L. (2012). Integration of three visualization methods based on co-word analysis. *Scientometrics*, 90, 659–673. <https://doi.org/10.1007/s11192-011-0541-4>

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.