Report for KING project.

A roadmap for mapping Food Science and Technology from scholarly publications on pulses.

Tristan Salord[a], Marie-Benoît Magrini[a*], Guillaume Cabanac[b], Marie-Josephe Amiot-Carlin[c], Marc Anton[d], Adeline Boire[d], Jean-Michel Chardigny[e], Colette Larré[d], Valérie Micard[f], Christophe Nguyen-Thé[g], Stéphane Walrand[h]

[a] AGIR, INRAE, Université de Toulouse, 31326 Castanet-Tolosan, France – tristan.salord@inrae.fr , marie-benoit.magrini@inrae.fr

[b] IRIT UMR 5505 CNRS, Université Paul Sabatier – Toulouse 3, 31062 Toulouse, France – guillaume.cabanac@univ-tlse3.fr

[c] MOISA, CIRAD, CIHEAM-IAAM, INRAE, Montpellier SupAgro, Université Montpellier, 34060 Montpellier, France– marie-josephe.amiot-carlin@inrae.fr

[d] BIA, INRAE, 44000 Nantes, France – marc.anton@inrae.fr ; adeline.boire@inrae.fr ; colette.larre@inae.fr

[e] DPTI, INRAE, Paris, France - jean-michel.chardigny@inrae.fr

[f] IATE, INRAE, Montpellier SupAgro, CIRAD, Université Montpellier, 34060 Montpellier, France – valerie.micard@supagro.fr

[g] SQPOV, INRAE, 84140, Avignon, France – christophe.nguyen-the@inrae.fr

[h] INRAE, UNH, Unité de Nutrition Humaine, CRNH Auvergne, Université Clermont Auvergne, 63001, Clermont-Ferrand, France – stephane.walrand@inrae.fr

[*]Corresponding author

E-mail: marie-benoit.magrini@inrae.fr (MBM)

Abstract

Mapping the evolution of a research field informs on its structure and its knowledge on which future research and innovations could be developed. Providing such analysis is essential to inform Science and Innovation policies, particularly for fields that need supports to be developed for the sustainability. Such a mapping can be achieved by combining bibliometric and text-mining approaches. Yet, no standardized method exists to conduct a large-scale analysis of a research field over decades, as opposed to classic reviews focusing on narrowly delineated topics. We defined a roadmap for mapping the research field of Food Science and Technology (FS&T) on pulses. This field is of particular interest, particularly in Europe, as pulse consumption needs to be developed for more sustainable and healthy diets. The scholarly output in this field keeps increasing, yet no study established a map of it to date. To initiate such a work, we queried the Web of Knowledge to delineate a corpus of 2,000 scholarly publications published between 1980 and 2018 on various pulse species and including at least one European author. Our approach combined the skills of experts in FS&T, scientometrics, and innovation economics. The clustering method used detected related research topics, revealing the structure of this field over time and the decelerating or front-run topics. Our analysis provides consistent information to Science and Innovation policies aimed at the development of pulses. Our developments also contribute to enrich the thesaurus for regenerating such science mapping on pulses, for new decades and/or by including a wider geographic area. Beyond this field, this methodological roadmap can be used for mapping any other research field.

Introduction

Mapping the evolution of a research field informs on the knowledge on which future research and innovations could be developed. A large consensus exists in STS (Science and Technology Studies) that Science and Innovation are strongly tied together [1]. Generally speaking, in all fields, expectations and prospects of firms (thereby defining their innovation paths) strongly depend on scientific advances [2]. Consequently, mapping the scientific research of a field is essential albeit challenging. Such a mapping can be achieved by combining bibliometric and text-mining approaches. Yet, no standardized method exists to conduct a large-scale analysis of a field of research over decades, as opposed to classic reviews focused on narrowly delineated topics. This article proposes a methodological roadmap applied to the research field of Food Science and Technology (FS&T) on pulses. This field is of particular interest as pulse consumption needs to be developed for more sustainable and healthy diets, particularly in Europe.

Indeed, legumes are an important asset in the sustainability transition of the European agrifood system [3-4], where both their cultivation and consumption are very low, accounting for only 2% of field crops (Table A in Appendix) [5-7]. Pulses are defined as a subgroup of legume plants from the Fabaceae (Leguminosae) family. According to FAO (Food and Agriculture Organization of the United Nations) pulses are annual crops used both for food and feed, yielding from one to several grains of variable size, shape, color with a pod. Since the International Year of Pulses in 2016 (the United Nations resolution A/RES/68/231), the European Commission has been giving more importance to food-based pulses through its new strategy for plant-based protein development [8]. Moreover, many studies recommend a significant increase in regular pulse consumption to achieve healthy effects in diets [9-10] compared with the current European consumption (4 kilograms/year/person, Eurostats). Poux and Aubert [11] proposed to reach 11 kilograms/year/person in Europe, whereas Willett et al. [4] suggested a consumption of 18 kilograms/year/person was appropriate for an universal healthy diet. Such an increased consumption will consequently generate higher pulse cultivation, favouring both sustainable and healthy diets [5].

However, pulses are facing a lock-in situation [12-15]. Since pulse consumption has been very low for years, suffering from an old-fashioned image and low consumption habits [16-17], weakened by low R&D investment [18], strong innovations are expected to reverse this situation and reach the recommended consumption levels. Moreover, even though new market opportunities emerge in pulse development, mainly driven by a trend towards reducing animal-based consumption, pulses are up against other alternative sources such as seaweeds, algae, insects, and cultured meat [19-21]. Above all, pulses compete with current major crops such as wheat or soya, which are still the main ingredients used in plant-based protein food innovation, as advanced in research studies [22-24] or market studies [25-26]. Therefore, even though some studies claim that "*it is clear that pulse proteins will be of utmost importance in satisfying food protein demands*" [27:p66] or that pulses require less high-technological processing than other alternatives [28], it is clear that this path-dependency means a shift in favour of pulses is still uncertain and will be difficult to achieve [7-15].

In front of this, scientific knowledge is essential for sustaining food innovation on pulses. In particular, the Food Science & Technology (FS&T) field on pulses is one of the keys to unlocking the situation. Therefore, informing on this field is essential. However, a global overview of this scientific area is lacking. In addition, based on a global analysis of the research output (scholarly publications indexed by the Web of Science) on both pulses and soya from 1980 to 2018, Magrini et al. [29] observed that research output for soya is twice more important than for pulses in the FS&T field; confirming the general lower R&D investment on pulses compared to other crops [18]. This observation questions the way the FS&T field develops scientific knowledge on pulses. Several reviews of the literature pointed out the various and numerous technological and functional properties of pulses (such as water and fat absorption, solubility, gel and emulsifying activity, foaming, fermentation…), but also remaining compositional and technological problems (such as digestibility, allergenicity, antinutritional factors…) [28]. But, to the best of our knowledge, no study managed to map the whole FS&T field on pulses to provide a wider perspective of the scientific knowledge accumulated on pulses.

Hence, the main objective of our study is to provide, for the first time, a map of the structure and evolution of the FS&T field on pulses in Europe from 1980 to 2018 based on a bibliometric dataset of 2,000+ publication records retrieved from the Web of Knowledge (Clarivate Analytics-WoS) which is a leading platform used in scientometric/bibliometric works. This objective relies on the huge progress in scientometric methods in analysing sciences through bibliometric data gathered over the past two decades [30-31]. The development of scientometrics led most high-level research institutes to developing scientific research programs for analysing Science itself [32-33], and helping Science and Innovation policy-makers to define new strategies based on the trajectories observed [34]. Indeed, we believe that having a mapping of this field, revealing and quantifying its structure through time, will help practitioners and stakeholders to define more accurate Science and Innovation policies to sustain pulse development for sustainability transition. This will provide an enhanced legibility of the field and will favour a more profound understanding of the directions taken in pulse development.

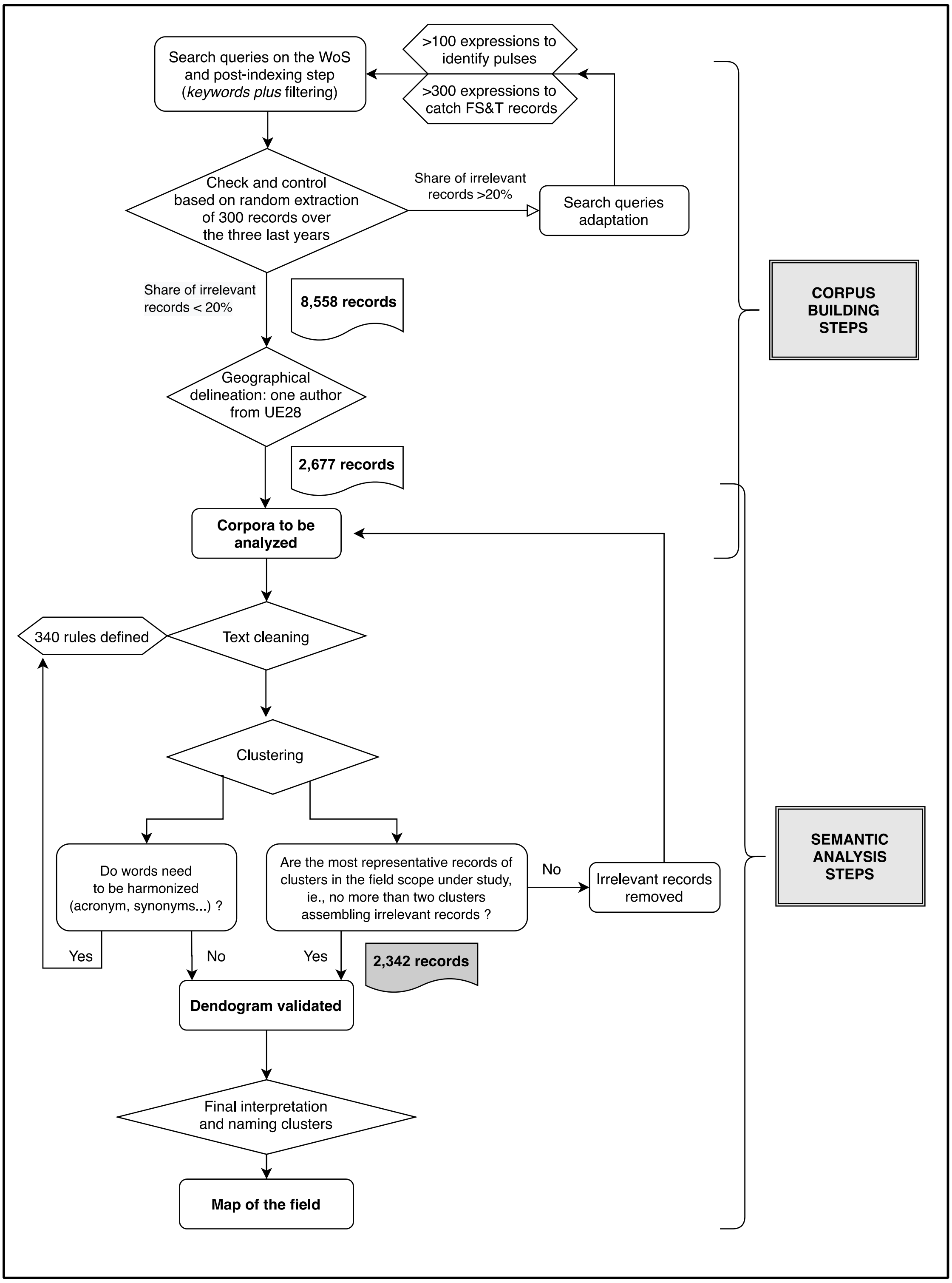
Considering the few previous works based on scientometric approaches in the FS&T field, two main strategies were deployed. Some authors used dedicated bibliometric platforms such as IFSA (Food Science and Technology Abstracts) [35] or bibliometric databases created by research institutes such as the Central Food Technological Research Institute of India [36].

Most authors built their corpora directly from large bibliometric platforms, such as Scopus or the Web of Science [37-41]. When large bibliometric platforms are used, filters and search queries have to be defined to delineate the FS&T field. Nevertheless, *(i)* it is still rare to find work that details these steps and makes the associated corpus available; *(ii)* another major issue is that most works were conducted by social or information scientists without the expertise of FS&T scientists to check the relevance of the corpora built [38-39;42]; *(iii)* these works mainly reveal the strategic position of countries in food science advancement or their relationships, but they fail to describe the ontology structure of the field; *(iv)* studies conducted by experts in the FS&T field reported (systematic) reviews of the literature on specific questions of interest, but did not provide a large bibliometric analysis encompassing the entire field. Hence, based on strong interdisciplinary work between experts from scientometrics, STS, and FS&T, our paper is a new step in the advancement of the analysis of the FS&T field. Beyond this field, our work provides an original methodological roadmap to map a research field over several decades and understand its structure and its scientific trajectories. Our approach is not limited to a classical study of the production of sciences based on descriptive statistics. We used text-mining and clustering methods to reveal the contents of the research field considered by identifying the main research themes that structure it, how these themes have evolved through time, and which ones appear more developed in the last decade.

Materials and Methods

This section explains the methodology followed as illustrated in Fig 1 and presents descriptive statistics on the corpus built to assess its relevance. Based on the work performed by Magrini et al. [29] that provided a bibliometric dataset from the WoS to analyse the evolution of scientific research (scholarly publications) on pulses and soya on a global scale in several scientific fields of importance (such as genetics, agronomy, animal nutrition, social sciences, and FS&T), our interdisciplinary consortium delineated from this bibliometric dataset (articles, book, book chapters, and reviews) a relevant corpus in the FS&T field of 2,342 scholarly publications between 1980 and 2018, on various pulse species (peas, dried beans, fababeans, lentils, chickpeas, etc.). All these publications include at least one European author. This final dataset is available from a public repository [43].

Fig. 1. Main steps to mapping a scientific field from a bibliometric dataset: the case of the FS&T field on pulses



Delineation of a corpus in the FS&T field on pulses

The starting dataset was based on search queries provided in [43]. This first set was established by ten experts from the FS&T field who delineated various FS&T subjects concerning mainly human nutrition, including health issues, allergies, processing and types of food uses (excluding non-food uses), sensorial and organoleptic analysis for consumer acceptance and dealing with pulses. Furthermore, as literature reviews on pulses are sparse and dedicated to the most commonly eaten pulse (such as the one conducted in [28] on peas, common beans, cowpea, fava beans, chickpea, lentil, or lupin), we considered 18 species and varieties of pulses cultivated in temperate climates (based on the taxonomy of [44] in order to broaden their analysis. Table 1 reports the various Latin and common names used to designate them.

Table 1. Expressions to identify pulses species and varieties cultivated in a European climate

Species or varieties identifier name	**All common or Latin names associated**
Adzuki	phaseolus angularis, vigna angularis, red mung$[1], red bean$, red mungbean$, adzuki$, azuki$
Bambara Bean	vigna subterranean *, bambara bean$
Bean	phaseolus coccineus, phaseolus vulgaris, phaseolus lunatus, phaseolus spp, common bean$, common field bean$, common fieldbean$, runner bean$, runnerbean$, lima bean$, common bean$, kidney bean$, pinto bean$, vigna aconitifolia, moth bean$, vigna umbellata, rice bean$
Chickpea	cicer arietinum, chickpea$, chick pea$
Cowpea	vigna unguiculata, cowpea$, cow pea, cow peas, blackeyed pea, blackeyed peas, black-eye pea, black-eye peas, blackeyed bean$, catjan$, long bean$
Faba bean	vicia faba, fava bean$, faba bean$, broadbean$, broad bean$, horse bean$, horsebean$, fababean$, field bean$, fieldbean$

1

Fenugreek	trigonella foenum grecum, trigonella foenum graecum, fenugreek$, fenugrec$, fenu grec$
Lathyrus	lathyrus sativus, lathyrus sativa, lathyrus ochrus, lathyrus cicera, grass pea$, red pea$, cyprus vetch$, vetchling$, gesse$
Gram bean	vigna mungo, gram bean$, black bean$, black lentil$, black gram, blackgram$
Lablab	lablab purpureus, hyacinth bean$, lablab bean$, lablab$
Lentil	lens culinaris, lentil$
Lupin	lupinus albus, lupinus angustifolius, lupinus luteus, lupinus mutabilis, lupin$
Mungbean	vigna radiata, vigna mungo, mungbean$, mung bean$, moong bean$, mungo bean$, green gram$, golden gram$, maash$, moong sanskrit$
Pea	pisum sativum, pea, peas
Pigeon Pea	cajanus cajan, pigeon pea, pigeon peas, pigeonpea$
Vetches	vetch$, vetche$, vicia sativa, vicia villosa, vicia ervilia, ervil$, vicia narbonensis, narbon bean$
Winged bean	psophocarpus tetragonolobus, winged bean$, asparagus pea$, goabean$, goa bean$

Note: $ (respectively, *) is a wildcard replacing zero or one character (respectively, no character at all or a group of characters) in search query.

Secondly, when retrieving a bibliometric dataset from the WoS, the “Topic Search” field collects records that match the user’s query on the following criteria: title, abstract, author keywords, and “KeyWords Plus”. But the “Keywords Plus” automatically added by the WoS to each bibliographic record lead to biased corpora. To circumvent this bias, we designed an indexing algorithm that cleaned the WoS results by filtering out documents with no direct interest (around 25% of the corpus upload from the WoS and as explained in [29]). Hence, our starting corpus delineation accounted for 8,558 records (articles, books, books chapters or reviews) published between 1980 and 2018 on pulses within the field of FS&T. In addition, we filtered out the records keeping those written by at least one author located in a European country (from UE28). Two main reasons led to this choice: the FS&T experts mobilized were strongly implicated at European level and so, their knowledge on pulses research from other regions in the world could be limited. This study aimed at informing the European R&D stakeholders, as pulses are under a new interest for the sustainability of European agrofood system as explained in the introduction. This geographical delineation decreased the corpus to 2,677 records.

Preprocessing steps

The semantic analysis targeted is based on a clustering analysis of words and sentences used in the titles/abstracts of records (explained in section 2.5). As a result, the relevance of the clustering analysis is strongly determined by the steps run to pre-process those words before analysis. Undoubtedly, in order to delineate relevant semantic topics, we have to reduce irrelevant terms, text ambiguities, synonymy and typographical errors, as much as possible [45]. Several steps were performed using the Iramuteq software (http://www.iramuteq.org).

Firstly, common expressions of no interest were removed (such as deictics or common terms such as “analysis”, “report”, “result”, etc.) and the remaining expressions were lemmatized: for instance, plurals were reduced to their singular form and conjugated verbs to their infinitive form.

Secondly, a normalisation step tackled the expressions of interest misidentified by this software: species synonyms were changed to their most common use (Species identifier name from Table 1), composed technical expressions were reduced to a common acronym (for instance, “high performance liquid chromatography” became “HPLC”). A total of 340 rules were defined with FS&T experts.

Thirdly, by conducting clustering analysis (Descending Hierarchical Clustering - CHD), we identified remaining records out-of-scope. They were identified through: *(i)* the semantic clusters that gathered terms not relevant to FS&T; *(ii)* reading abstracts from the most representative records contributing to relevant clusters to check if some records contained words that had a different meaning to those in FS&T.

In our case, three clustering steps were generated. We stopped when only two (close) clusters gathered irrelevant records while most of records contributing to other clusters were relevant to the field. As we explain in Section 3, the final dendrogram we kept (Fig 8) presents two clusters (Clusters 11 & 2) gathering irrelevant records. Those "out of scope" clusters mainly dealt with the subject of pulse protein characterisation or metabolism in plant science studies.

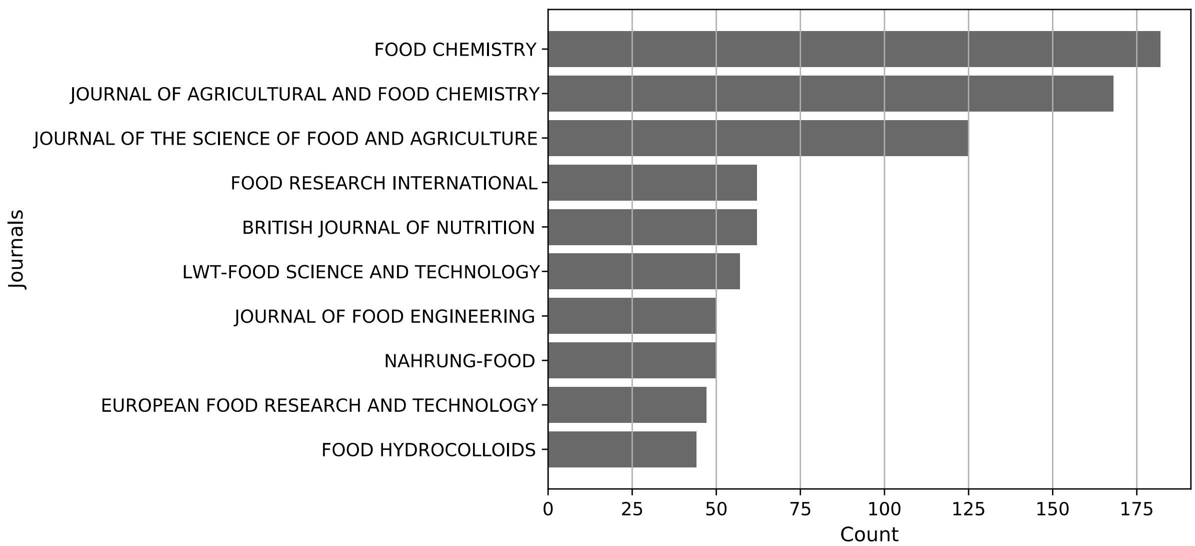
After these cleaning steps, the corpus accounted for 2,342 records, focusing on the European scientific FS&T field on pulses over the last four decades: we called it the "European FS&T pulses corpus".

Representativeness of the corpus

In order to assess the representativeness of the corpus analysed, we provided some statistics on the most frequent journals and WoS categories of the records.

Firstly, these records were published in 586 journals. The "Top 10 journals" gather 36% of the records published (Fig 2) whose journal titles confirm the representativeness of the corpus, and 60 journals concentrate 75% of the records.

Fig. 2. The 10 most frequent journals associated with records in the European FS&T pulses corpus.



Secondly, the WoS categories of journals (Fig 3) also confirm the relevance of the corpus through various specialised categories related to food. We observed that the FS&T WoS category dominates over time, notably due to the greater specialisation of journals over time (Fig 4); however, a significant proportion of records are published in journals categorised differently by the WoS. Considering a larger array of WoS categories to map a research field would allow to collect papers with interdisciplinary approaches.

Fig. 3. The 10 main frequent WoS categories of journals in the European FS&T pulses corpus (fractionated counts used when the WoS attributed several categories to the same journal).

[INSERT FIG 4]

Fig. 4. The longitudinal evolution of the 10 main frequent WoS categories (WC) of journals in the European FS&T pulses corpus.

Thirdly, note that the CAGR (Compound Annual Growth Rate) of records is higher over the period 1990-2000 (Table 2) because before 1991, only titles were indexed by the platform. Considering the previous or current decades, the growth of records from the FS&T pulse field is just above the annual growth rate in scholarly peer-reviewed English-language journals of the WoS that increased 5–6% in recent decades [46]. This last point means that the quantity of studies in the FS&T field on pulses is no higher than any other scientific fields.

Table 2. Number of records by periods and compound annual growth rate (CAGR).

Period	**CAGR**	**Number of records**
1990-2000	7%	624
2000-2010	4%	742
2010-2018	3%	976

The geographical dimension of the corpus

We computed the shares of records among countries in UE28 and found a country-ranking close to the ranking from other studies in FS&T even if it is difficult to compare rankings from one study to another as different counting methods and/or bibliometric platforms can be used. Nonetheless, the ranking we established, based on a fractional counting for records associating several countries (each country accounts for 1/n where n is the number of countries) is close to other rankings in FS&T from other studies which found on the whole field of FS&T over the period 2003-2013 (from the Scopus platform) that Spain, Germany, Italy, UK and France to be the major publisher countries [39]; or on the WoS for a larger field on agrifood research that found to be UK, Germany, France, Spain, Netherlands [38]. Moreover, our ranking of countries changes marginally, considering a fractional counting or a whole counting relative to the location of the first author (Fig 6). This ranking appears stable with a dominant position of 7 countries over the last decade: Spain, Italy, Germany, Poland, France, United-Kingdom, and The Netherlands. We also calculated the ranking based on the citation count in 2018 (Fig 7), which corresponds to the number of citations accumulated by records up to 2018 (fractionated count for records associating several countries). Whatever the ranking method, Spain remains the foremost publisher on pulses in the FS&T field (being also the European country with the highest pulses production).

[INSERT FIG 5]

Fig. 5. Ranking of the European countries over the 1980-2018 period by record count (fractional counting on the left, first author counting on the right).

[INSERT FIG 6]

Fig. 6. Ranking of the European countries over the 1980-2018 period by citation count (fractional counting on the left, cumulative counting on the right).

Methodological choices for the semantic analysis based on clustering methods

Numerous algorithms (like LDA, LSM, CTM, knowledge-based graph approaches,…) aim to discover and extract semantic (or thematic) subsets within texts. These methods generally use "Natural Language Processing" [47-48] or "Text Mining" methods. These algorithms focus on extracting topics from a set of unstructured textual data, yet these algorithms cannot be used one for another because each one corresponds to a particular idea of what a topic could be delineated. The choice of using one method/algorithm is a compromise between the type of data studied, the objectives, and the search routines of the researcher(s).

In other words, the choice of one method is not done in a deterministic way. For example, a popular algorithm for topic extraction, LDA, faced some criticisms concerning the way topics are conceptualized and the method used to reconstitute them. Shadrova [49] detailed the issues this method had to face; and particularly underlined the fact that this algorithm is a heuristic (i.e., stochastic) method, not a deterministic one, which means that different topics could be extracted from the same textual data for a same number of topics. This greatly hinders the reproducibility of the results, and explains one of the reasons why we did not choose such an algorithm. Another major reason relies on the fact that such an approach, and more generally, algorithms based on the same way textual data are grouped (a method called "bag-of-words"), hardly take account of complex multi terms expressions, such as we used to use in science, as the citation context of words.

On the opposite, the Reinert's method [50], implemented in the Iramuteq software we chose to use, is based on taking into account the contexts of citations, as Sbalchiero [51] summarized it *"from a theoretical standpoint (...) it is just not a matter of studying occurrences and co-occurences of words in a text, but of understanding their relationships in the discursive context".* This method has also the advantage of being deterministic: its results do not change, regardless of the number of times it is executed on the same corpus with the same settings given as inputs.

As for many other algorithms dedicated to topics extraction, the method is initiated based on a the user's choice of the number of topics to discover – we will explain later on how we manage this parameter. Once initiated, the algorithm tends to identify as many thematic clusters – what Reinert calls "lexical worlds"- in texts as it was determined by the user, on the ground of their lexicographic proximity. This proximity is computed based on the analysis of co-occurrence of words between what Reinert calls Elementary Context Units (ECU) or "Text Segments", which are parts of a sentence or a whole sentence, extracted from the title/abstract of records. As already stressed, one of the most pertinent advantages of this method is that the words analysed are not isolated from their context of appearance.

The classification algorithm executes a series of bi-partitions using Factorial Correspondence Analysis (FCA) on text segments. Hence, the FCA is based on a matrix crossing text segments and words. The resulting output consists in a dendrogram produced by a Descending Hierarchical Clustering (DHC) ran iteratively that distinguishes classes of text segments whose vocabulary differs the most, and tends to produce as many clusters as it was asked at the beginning of the process [50; [51] 52-53]. Applied to our corpus, the algorithm distinguished 18,783 distinct words from 10,371 text segments representing 87% of all text segments in the corpus, and concerning 2,342 bibliographic records (meaning that only 44 records present no text segments matching those from other records). This lexicon of 18,783 words accounts for 366,158 occurrences.

As we mentioned it, one major issue of topic extraction algorithms concern the choice of the number of topics to be identified. In most cases, this parameter is left to the user's choice and no statistical method can help presume *a priori* the right number of topics a corpus contains.

To get around this issue, we recourse to a panel of 9 experts from various subfields in FS&T that helped stabilize the number of topics to consider by investigating the DHC.

The workflow was the following. We initially produced a lexical analysis looking for the largest number of topics we could obtained. Experts were collectively asked to name each node and cluster of the DHC produced. If they reported difficulties of understanding during the process or were not able to recognize a unity of meaning for each cluster, we reduced the target number of clusters, producing another DHC. At each step of the process, the fifteen most representative publication records of each cluster (making it possible, as mentioned above, to check the relevance of the records contributing to the cluster under analysis) and the most frequent text segments from each cluster were provided to the experts. In order to reinforce the interpretation of the topic delineation of clusters and to help experts labelling and describing each cluster, we also calculated correlations of specific terms with clusters. We chose terms frequently used by experts during work sessions, and determined in which clusters those terms were over-represented compared to other clusters. This process was iteratively conducted until we reached a clustering correctly labeled. Five iterations were necessary to produce the results presented in this paper.

The temporal evolution of those clusters was also considered, allowing us to determine which subjects/topics were already strongly present in the beginning of the period; and which ones were more investigated in recent years, then considered as forefront topics. This analysis was based on a chrono-dendrogram representing the temporal projection (by year) of the clusters' evolution. It was calculated in terms of text segment proportions over time. Based on the Chi^2 test, we also identified the yearly over-representativeness of clusters. One must consider the longitudinal results with caution because the editorial coverage of the WoS has increased during the period; before 1991, only titles were indexed by the platform. These results are presented and discussed in the following section.

Results and discussion

The core of our analysis relies on semantic clusters to establish an original map of the FS&T field on pulses in Europe over the last decades. We comment them via five main types of figures generated from the methodology explained in the last section:

1. The dendrogram: Fig 8 is the final dendrogram composed of 11 clusters interpreted by experts. The figure reports the 20 most representative words of each cluster, and the number of text segments (TS) delineated by the software to establish each cluster. The figure reports the titles attributed by experts to the main subfields structuring this dendrogram, and a short description of each cluster according to their main focus (based on the analysis of the TS that most contributes to clusters). In the following discussion of clusters, some excerpts taken from records are reported to illustrate the meaning of the words identified in clusters.
2. The correlation tests of selected terms with clusters: they are introduced along the discussion to reinforce the interpretation of clusters.
3. The chrono-dendrogram: Fig 9 represents the evolution of the clusters over time.
4. The Chi^2-chrono-dendrogram: Fig 10 represents the clusters over-represented by years. It allowed us to identify the forefront topics.
5. Table 3 synthesizes information from the aforementioned figures to establish a ranking of clusters regarding their importance in recent years.

Based on these results, sections 3.1 to 3.4 describe the various topics studied in the field of FS&T on pulses and their evolution. The last section (3.5) discusses the main trends emerging from this analysis to identify innovation paths on pulses.

[INSERT FIG 7]

Fig. 7. Final dendrogram from the European FS&T pulses corpus (1980-2018). TS: text segments.

[INSERT FIG 8]

Fig. 8. Chrono-dendrogram from the European FS&T pulses corpus (1980-2018)

[INSERT FIG 9]

Fig. 9. Chi2-chrono-dendrogram from the European FS&T pulses corpus (1980-2018)

Table 3. Ranking of Clusters based on their over-representation over the period 2014-2018.

Cluster number	**Nb. records (1980-2018)**	**Nb. records (2014-2018)**	**Share of records from 2014-2018 per.**	**Nb. text segments (1980-2018)**	**Nb. text segments (2014-2018)**	**Share of text segments from 2014-2018 per.**	**Nb. years with over-representation of the cluster between 2014 and 2018**
6	243	104	43%	700	305	44%	**4**
1	256	123	48%	734	394	54%	**3**
4	366	162	44%	540	263	49%	**3**
7	482	175	36%	1003	375	37%	**2**
3	205	80	39%	596	260	44%	**2**
8	251	92	37%	572	206	36%	**2**
11	533	124	23%	1291	258	20%	**0**
5	347	76	22%	1181	216	18%	**0**
2	562	141	25%	1176	282	24%	**0**
9	500	155	31%	887	261	29%	**0**
10	326	84	26%	581	131	23%	**0**

The over-representation is established by a Chi^2 test on the text segments proportion (p-value < 0.01). Consequently, even if a cluster presents a lower number of records compared to another cluster, it could be overrepresented in recent years if numerous TS (text segments) contribute to the cluster over recent years compared to an earlier period.

Main structure and evolution of the European FS&T field on pulses over 1980-2018

Firstly, in Fig. 9 and 10, we observe a swing in time of two main types of works on pulses: over the first decades, there were more studies on "Molecular characterization of proteins" than those focusing on "Food applications, health & nutritional effects", which is no longer the case since the end of the nineties.

Secondly, we observe an increase in activity on the subjects of "Nutrition, Health and Consumers", as more and more research protocols are based on humans and less on lab animals; conducting also a recent higher focus on food acceptance and consumers' education.

Thirdly, inside the subfield "Food Applications, health & nutritional effects" two main focuses appear: *(i)* fraction processing (mainly starches and proteins, proteins being more and more investigated in recent years); *(ii)* digestibility issues with an earlier focus on the control of ANF (anti-nutritional factors), whilst nowadays there are more studies on other nutrients such as vitamins and minerals.

As mentioned in Section 2, we stopped clustering steps when no more than one or two clusters gathered irrelevant records whilst most of others were constituted of relevant records for the FS&T field. In Fig. 9, the "out of scope" clusters concern Clusters 11 and 2 on the "molecular characterization of proteins". Nevertheless, while the records from those clusters might not directly concern the scope of FS&T, they do bear interesting knowledge as regards the characteristics of proteins in pulse seeds. That is why we kept them in the final dendrogram, and that the software effectively identified both their proximity (appearing close to each other) and differences from other clusters (appearing at one extreme side of the dendrogram). Most records from those clusters were published in the early stages, and frequently in plant science journals. Cluster 11 is dedicated to the study of metabolic activities of proteins in plants at a molecular scale (words such as "molecular", "molecule", "genetic" are the most frequently correlated to Cluster 11 compared with all other clusters); whilst Cluster 2 focuses mainly on the study of storage activity of pulse proteins at the seed scale. Not being in the core scope of the FS&T field, we will not comment further on those two clusters in the rest of this article.

Clusters focusing on health and nutritional effects

Records from Clusters 5 and 6 (located at the right of the dendrogram) are related to specific public health issues on food intake in western countries, a growing subject. The split between these two clusters mainly refers to the research protocol being with human beings or laboratory animals. These records share similar issues on health and nutrition. However, Cluster 6 presents a broader scope with studies on consumers, and mainly on healthy consumers. These clusters tend to follow each other in time: Cluster 5 gathers earlier records compared with Cluster 6, over-represented in recent years. Half of the records contributing to Cluster 6 were published between 2014 and 2018, whilst only 18% for Cluster 5. Hence, this shift reveals that much research in the FS&T field relied initially on animal laboratory studies, before dealing with human beings.

We observe that a central topic for these two clusters concerns weight and obesity issues (improvement of satiety, gut function, reduction of lipid accumulation, etc.), weather researches are based on animal studies or human studies, and as the excerpts below illustrate. In addition, we observe that the terms "obesity" or "satiety" are over-represented in those clusters (Fig 11) with a strong focus on appetite sensation in Cluster 6.

> Cluster 5 - WOSID:000072370100013 - "Lipid accumulation in obese Zucker rats is reduced by inclusion of raw kidney bean (Phaseolus vulgaris) in the diet.(...) results suggest that, in addition to animal nutrition, it may also be possible to use the bean lectin as a dietary adjunct or therapeutic agent to stimulate gut function and ameliorate obesity if a safe and effective dose-range can be established for human subjects."
>
> Cluster 5 - WOSID:000376881700034 - "Effects of Dietary Fibre (Pectin) and/or Increased Protein (Casein or Pea) on Satiety, Body Weight, Adiposity and Caecal Fermentation in High Fat Diet-Induced Obese Rats.(...) Altogether these data indicate that high fibre may be better than high protein for weight (fat) loss in obesity."
>
> Cluster 6 - WOSID:000396880100001 - "Meals based on vegetable protein sources (beans and peas) are more satiating than meals based on animal protein sources (veal and pork) - a randomized cross-over meal test study. (...) Interestingly, a vegetable-based meal with low protein content was as satiating and palatable as an animal-based meal with high protein content."

Cluster 6 - WOSID:000240728800005 - "Efficacy of 12 weeks supplementation of a botanical extract-based weight loss formula on body weight, body composition and blood chemistry in healthy, overweight subjects - A randomised double-blind placebo-controlled clinical trial.(...) extracts of Kidney bean (...) A significant change of the Body Composition Improvement Index and the decrease in body fat was statistically significant in active extract subjects compared to placebo."

[INSERT FIG 10]

Fig. 10. Clusters correlation with "obesity, satiety" lexical field (Chi2 test).

Whereas Cluster 5 and 6 share common research issues, Cluster 6 concerns a broader range of topics as more pulse compounds (such as fibre, lectin, flavonoids, phytates, phytoestrogens, …) are mentioned within the records. In particular, pulse-based fibre issues are mainly considered as a means to improve appetite sensation in meals that could be meat-replacement meals or meat-based enriched with plant proteins. On the whole, Cluster 6 assembles more meal or diet-based studies compared to all the other Clusters.

Cluster 6 - WOSID:000318710200001 - "Self-Reported Food-Related Gastrointestinal Symptoms in IBS Are Common and Associated With More Severe Symptoms and Reduced Quality of Life.(...) Symptoms related to intake of food items with incompletely absorbed carbohydrates were noted in 138 (70%) patients; the most common were dairy products (49%), beans/lentils (36%), apple (28%), flour (24%), and plum (23%)."

Cluster 6 - WOSID:000378275800005 - "Protective Effect of Myo-Inositol Hexaphosphate (Phytate) on Abdominal Aortic Calcification in Patients With Chronic Kidney Disease.(...) Among the top 10 phytate-rich foods, lentil consumption was significantly greater in patients with no/mild AAC than in those with moderate/severe AAC."

Cluster 6 - WOSID:000411807800014 - "We investigated whether ad libitum energy intake (EI), appetite, and metabolic markers in a meal context were affected by 1) fibre addition (rye bran and pea fibre) to pork meatballs, 2) the food matrix of the fibre (fibre meatballs compared with fibre bread), or 3) the protein source (animal compared with vegetable protein patties)."

We observe an emergent topic of interest in Cluster 6 concerning notably food education and food acceptance issues, revealing a broader scope of challenges in fostering pulses in diets. Even though none of the studies on food education are actually dedicated to pulses, some studies focusing on vegetables did include pulses (most often peas, beans or lentils) and these studies are recent ones, which gives us primary insights on these issues.

> Cluster 6 - WOSID:000080857500005 - “Repertory-grid method obtained information from the children regarding their perceptions of eight vegetables commonly consumed in the UK (baked beans, carrots, tomatoes, cauliflower, turnip (swede), cabbage, sweetcorn and peas)”
>
> Cluster 6 - WOSID:000348270100026- “the impact of food neophobia in food habits and preferences of healthy food in school can- teens users in the city of Murcia (...) A higher level of neophobia less acceptance was given to foods like chicken and lentils ($p<0.05$), fruit, salads and legumes ($p<0.001$)”

Clusters focusing on food applications

These different clusters cover two main subfields: one concerns the processing and techno-functional properties of pulse fractions (Clusters 1, 7, 8 3) with a recent focus on “Enriched wheat-based products” (Cluster 1), and another, nutrition issues mainly linked to digestibility concerns (Clusters 4, 9, 10) with a recent focus on new promises for consumers (Cluster 4).

Clusters 1, 7, 8, and 3, focusing on the food applications with fractions from pulses, assemble more recent works (the proportion of records published in recent years is respectively 48%, 36%, 37%, and 39%) compared to Clusters 9 and 10.

The clusters from the first subfield (Clusters 1, 7, 8, 3) are divided according to the type of fractions studied: Clusters 7 and 8 focus specifically on pulse starches, whereas Cluster 3 focuses more on proteins (and assembles numerous records in a recent period). More precisely, issues in Clusters 7 and 8 are directed to starch extraction methods (dry and wet milling), their functional and rheological properties and the benefits of their use in industry. Cluster 7 refers more to specific methods, notably a number of TS refer to advanced computational methods requiring artificial intelligence (deep learning methods, neural networks, etc.). Cluster 8 reflects publication records that mainly address the functional and rheological properties of pulse starches and modified pulse starches (note the presence of the “annealing method” in the cluster). They especially focus on the different steps of the gelatinization process and their consequences on the structure of pastes using such starches (the structure of amylose and amylopectin of pulse starches, the rheological of different starches during the heating process…). This cluster particularly refers to the review from the literature by Hoover on starches [54].

Cluster 3 addresses a topic of growing interest: proteins, with a particular focus on concentrate and hydrolysates. Compared to previous clusters on starches that reflected a high number of publication records, Cluster 3's records are fewer in number but over-represented in recent years. The issues with proteins concern both ways to decrease production costs, to improve processing methods, to reduce allergenic risk or to add nutraceutical advantages to products. Part of these records concerns notably the solubility, foaming, interfacial and gelling properties of proteins as are shown in the TS extracts below.

> Cluster 3 - WOSID:000373168200022 - "Functional properties of pea protein hydrolysates in emulsions and spray-dried microcapsules. (...)This effect may be attributed to the altered physical properties of the interfacial film as well as the antioxidative effects of the hydrolysed proteins."
>
> Cluster 3 - WOSID:000291917000012 - "Foaming activity of lupin protein isolates in the absence of insoluble protein aggregates. This study investigates the foaming properties of two lupin protein isolates (LPIs) extracted by ultrafiltration (albumin-rich fraction; LPTF) and isoelectric precipitation (globulin-rich fraction; LPTE) (...)."

Cluster 3 also gathers publication records focusing on new techniques regarding the uses of pulse proteins (physical, enzymatic methods, …) for improving their physicochemical properties.

> Cluster 3 - WOSID:000373168200022 - "Functional properties of pea protein hydrolysates in emulsions and spray-dried microcapsules. In the present study the impact of partial enzymatic hydrolysis on the functional properties of pea protein isolate (PPI) was investigated."

All in all, issues in Cluster 3 present strong similarities with the review of the literature of Bessada [28] on pulses proteins whose interest is rising because of the various functional properties associated with pulses.

Whereas Clusters 7, 8 and 3 mainly deal with the processing and functional properties of pulse fractions, Cluster 1 groups together text segments strongly related to a specific sector of the agri-food industry: enriched wheat-based products (such as bread, bakery, pastry or dough products). The correlation terms (Fig 12) confirm this, as the terms “bread”, “bake”, “bakery”, “crust”, “crumb”, “crumbgrain”, “crumbliness” are over-represented in Cluster 1. It is a sector of rising interest, confirmed by a recent review on this topic [55].

[INSERT FIG 11]

Fig. 11. Clusters correlation with “bread, bakery and pastry” lexical field (Chi2 test).

In particular, issues in records from Cluster 1 concern the effects induced by adjunctions of various pulse flours and compounds for improving rheological, sensorial, organoleptic or physicochemical properties of bread - and less frequently cakes - doughs. Hence, pulses adjunction appears to be of high interest for bread-making - improving nutrient quality, useful for gluten-free products - but impacting the physicochemical properties of dough, bread texture and colour as illustrated in the following TS.

Cluster 1 - WOSID: 000180850500010 - “Dough structure and bread quality [are] inversely related to the supplementation level of [pulses] flours”

Cluster 1 - WOSID:000343818700006 - “Effect of the lentil flour and hydrocolloids on baking characteristics of wholemeal durum wheat bread. (...) The nutritional analysis of the optimised functional bread showed an increase in proteins and dietary fibre (total, soluble and insoluble), suggesting its high functional value and the possibility to propose it on a market that is constantly changing and attentive to health benefit of foods.”

Cluster 1 - WOSID:000438137700045 - “Dextran produced in situ as a tool to improve the quality of wheat-faba bean composite bread. The incorporation of faba bean flour into wheat-based products is a sustainable way to obtain protein-enriched food items. However, developing breads with a higher content of faba bean flour is challenging due to the poor textural/sensory properties of the final product. A potential solution is to use hydrocolloids as structuring agents to increase the viscoelastic properties of the composite bread.”

Consequently, a substantial part of TS in Cluster 1 focuses on processes aiming to decrease the negative effects of pulse adjunctions in terms of textural and sensorial dough properties, notably through fermented or germinated pulses flours more currently investigated [55-56], and illustrated below in TS extracts.

> Cluster 1 - WOSID:000180850500010 - “Changes of wheat dough and bread quality and structure as a result of germinated pea flour addition. (...)Differences in microstructure of dough and bread were found to be closely related to the textural properties of bread.”
>
> Cluster 1 - WOSID:000387518800024 -” The influence of lactic acid fermentation on functional properties of narrow-leaved lupine protein as functional additive for higher value wheat bread. (...)The measurement reflected the introspective liking ratings, especially "happy" and "sad" showed a high correlation with liking and were good indicators for liked and disliked samples, respectively.”

Finally, Cluster 1 is also characterised by a higher number of records with acceptability studies dealing with organoleptic properties issues:

> Cluster 1 - WOSID:000404792200023- “The comparison of the effect of added amaranth, buckwheat, chickpea, corn, millet and quinoa flour on rice dough rheological characteristics, textural and sensory quality of bread. (...) Overall acceptability of composite breads containing amaranth, chickpea and quinoa was negatively impacted by the aroma and taste of these flours.(...)”
>
> Cluster 1 - WOSID:000403932300017 - “Waffle production: influence of batter ingredients on sticking of fresh egg waffles at baking plates-(...). This resulted in significant correlations with take-off-force, which was correlated with L*-and b*-value (negative) and positive to a*-value. Sticking behavior was strongly associated with b*-value (positive) and to a*-value (negative).”

Clusters of the second subfield (Clusters 4, 9 and 10) mainly focus on nutrition issues. Because pulses can be an alternative source for essential nutritional compounds (such as amino acids, minerals, antioxidants), they bear significant new *promises* as often advanced in TS assembled in Cluster 4. However, the bioavailability of such compounds of interest strongly depends on the way anti-nutritional factors in pulses (ANF such as trypsin inhibitor [57]), tannins, polyphenols, pythtates…) are degraded or deactivated by different process methods [58]. Clusters 9 and 10 (close to Cluster 4 in the dendrogram) detail such promises, by assembling records focusing either on the industrial opportunities offered by pulses, or on various factors (storage methods, process methods, seed maturity stage) impacting on the bioavailability and bioaccessibility of pulse compounds or by-products of interest (starches, amino acids, antioxidants, nutrients, minerals). However, Clusters 9 and 10 concern less recent records as the proportion of records published between 2014 and 2018 are respectively 31% and 26%.

Beyond the protein interest, what distinguishes clusters 9 and 10 are mainly the compounds studied. Cluster 10 is mainly centred on the bioavailability of minerals contained in pulses (such as zinc, iron, phosphorus, magnesium, calcium, etc.).

> Cluster 10 - WOSID:000180380500027 - "Effect of different soaking solutions on nutritive utilization of minerals (calcium, phosphorus, and magnesium) from cooked beans (Phaseolus vulgaris L.) in growing rats. (...) The increased amount of cellulose in the soaked seed did not have a negative effect on the digestive utilization of minerals."
>
> Cluster 10 - WOSID:000172704400005 - "In vitro extractability of calcium, iron, and zinc in finger millet and kidney beans during processing. (...) Addition of vitamin C and mango could be used to enhance mineral extractabilities, thereby helping to alleviate micronutrient deficiencies in populations subsisting on these foods."
>
> Cluster 10 - WOSID:000075599000010 - "Nutrients and antinutritional factors in faba beans as affected by processing. (...) Of all the treatments studied, germination appears to be the best processing method to obtain nutritive faba bean flour, since it caused a minor decrease in starch content (15% loss), the largest alpha-galactoside and phytic acid removal (94% and 45%, respectively) and provided an appreciable amount of dietary fibre."

Whereas Cluster 9 focuses on other pulses compounds issues, i.e the preservation of the antioxidant's properties of pulses, the conservation of amino acids and starch quality:

> Cluster 9 - WOSID:000185164500016 - "Antioxidant properties of lupin seed products. (...)Increasing doses of irradiation lowered antioxidant effects of lupin extracts; however, the antioxidant activities of some samples were higher. The observed negative changes in the tocopherols contents were effects of the irradiation dose as well as storage time."
>
> Cluster 9 - WOSID:000327169500001 - "Phenolic compound composition in immature seeds of fava bean (Vicia faba L.) varieties cultivated in Chile.(...) The results from this study revealed clear differences in the phenolic composition among different varieties of immature V. faba L seeds and demonstrates that there is ample phenotypic variability for future selections studies for traits such as nutritional value, taste, and ease of production."

The processing methods assessed in these clusters also diverge. Cluster 10 refers more to traditional processing such as soaking, germination and fermentation methods, in contrast to

Cluster 9 where methods are more related to industrial purposes. The over-representation of terms related to tradition (such as "tradition", "traditional", "traditionally") supports this interpretation (Fig 13).

[INSERT FIG 12]

Fig. 12. Clusters correlation with "tradition" lexical field (Chi^2 test)

Cluster 4 assembles new promises on pulse-based food products. These promises can be technological – the advantages in introducing specific pulse flour in pasta composition, reducing microbiological risks -, sensorial - improving palatability of specific foods, reducing off-flavour -, or nutraceutical, focusing on specific processing methods and/or specific pulse by-products, as illustrated by the TS below.

Excerpts on technological & sensory promises:

> Cluster 4 - WOSID:000395875400040 - "Influence of fermented faba bean flour on the nutritional, technological and sensory quality of fortified pasta. Faba bean has gained increasing attention from the food industry and the consumers mainly due to the quality of its protein fraction.(...) The use of fermentation technology appears to be a promising tool to enhance the quality of pasta and to promote the use of faba bean flour."
>
> Cluster 4 - WOSID:000394265600006 - "Flavor Aspects of Pulse Ingredients. Pulses (Fabaceae) have regained interest for their high protein level, However, food application of pulses and pulse ingredients is hampered by several issues around their off-flavor.(...)This review article aims to provide a concise overview highlighting the most important aspects of the knowledge available on the off-flavor compounds present in various pulses, their possible origins, and the technologies available to prevent, reduce, or mask these off-flavor compounds.(...)"

Excerpts on nutraceutical promises:

> Cluster 4 - WOSID:000394632000010 - "(...)The data obtained demonstrates that the agro-industrial by-product could potentially be utilized as a promising source of bioactive ingredients to design new functional foods and nutraceuticals with a valuable future market. Furthermore, the obtained data may form a basis for future quantitative and bioavailability studies, which will be the next step in this present work."
>
> Cluster 4 - WOSID:000331851400029 - "In vitro antioxidant activities and antioxidant enzyme activities in HepG2 cells and main active compounds of endophytic fungus from pigeon pea [Cajanus cajan (L.) Millsp.]. (...) The present investigation suggests that MD-R-1 represents a valuable natural antioxidant source and may be applicable in health food industry potentially."

Note also that these "new promises" are mainly mentioned through industrial food applications and not through studies that could have encompassed the whole diet composition. It is clear that the lexical field of industry is stronger in Cluster 4, distinguishing it from Clusters 9 and 10. The Fig 14 demonstrates the over-representation of words such as "industry", "industrial", "market", "benefit", "brand" in Cluster 4.

[INSERT FIG 13]

Fig. 13. Clusters correlation with "industrial" lexical field (Chi^2 test).

Therefore, it does not come as a surprise to observe that allergy issues are also over-represented in this cluster, as food industry allergies is clearly a subject of major interest. Also, the Fig 15 illustrates the over-representation of the lexical field of allergy (considering the terms "allergy", "allergic", "allergenic", "allergen", "allergenicity") in Cluster 4:

[INSERT FIG 14]

Fig. 14. Clusters correlation with "allergy" lexical field (Chi2 test).

Which structuration of FS&T for pulses?

This mapping and the time evolution of the semantic clusters reveals several main insights.

Firstly, it points out a general organisation of the FS&T field on pulses: the consolidation of knowledge at a molecular scale and on processing solutions are first required before investing more studies on food applications. In addition, these food applications cover both traditional and more technological applications - the latter are increasing. When works on food applications are increasing, this creates a path for more works advocating "new promises for consumers"; proposing new foodstuffs corresponding to societal expectations given the developments achieved in the other subfields. We observed that this semantic cluster around "promises for consumers" is based on recent publications: it tends to demonstrate how consumer expectations are more taken into account by the FS&T field. Nevertheless, food acceptance and education on pulse consumption still lack investigation, no doubt due to low consumption habits. This suggests that research on consumer behaviour (perception, belief, knowledge and barriers) needs to be reinforced. Consumer acceptance is vital for developing pulse-based food products and sustain pulse-chain value.

Secondly, another structuration of this field enlightened two main technological domains: one on pulse-based fractions and compound valorisation (for instance, protein extracts), another one on pulse whole-grain valorisation (for instance, flours). It would appear that the first path has received more investigation, revealing the "technologisation" of grains, confirming how food is seen more and more as a sum of ingredients, a combination of nutritional and technological properties. Nevertheless, even if fractionation processes are developing on pulses as they have been developed on soya or cereals, the valorisation of whole-grain pulse-based applications remains a significant part of research in the FS&T field. However, in a context where there is a clear increase in vegetarianism/flexitarianism, meat analogues or substitutes - which are strongly linked to the technologisation of food - are poorly investigated for pulses. Indeed, occurrences of expressions such as "meat analog", "meat substitute", "meat like" count for less than 10 in the whole corpus (among titles and abstracts). This confirms what we mentioned in the introduction: the market of meat replacement is currently more invested with major species than with pulses, particularly with soya, even if pea is often quoted as an interesting option [59]. The low availability of pulse production in agricultural markets prevents the food industry from further investing, and consequently industry prefers developing meat substitutes with soya as it is so largely cultivated throughout the world.

Thirdly, there is a clear tension between the studies highlighting the positive effects of pulses and those dealing with their negative effects. Undoubtedly, the beneficial effects of pulses are more and more highlighted regarding the technological (using pulse flours to improve rheological qualities of a dough), gustatory (adding texture or flavour effects to food preparations), and health-related aspects (e.g. diabetes prevention), whereas many other studies are still dealing with the control of ANF and sensoriality dimensions (e.g. dough properties, texture and palatability). This shows why more research works concerning the limits in using pulses are clearly still needed, and that Science and Innovation policy makers must wonder whether the allocated research means are sufficient to rapidly overcome these limits, or at least to allow pulses to break through in the food industry, and make the consumption of pulses more common in order to foster sustainable diets.

Fourthly, the use of pulses in bakery is a rising topic both for incorporating whole-grains (flours) or fractions of pulses. Those food applications could be strategic to increase the daily intake of pulses through commonly consumed foods (bread, pasta...). Whereas there is marked development in sensorial studies regarding organoleptic perceptions, we observed that few works dealt with eating habits for questioning food acceptance and education. These last issues are essential when expecting a better market position of future innovative bakery products. In addition, the beneficial contribution of innovative pulse products needs to be evaluated in the whole diet. Considering the "global health" challenge, for humans and the planet, food and technological research on pulses could be a selected target for Science and Innovation policy.

Food research and innovation on pulses

Results from this work suggest new directions for Science and Innovation policies on pulses.

Firstly, several topics appear as promising building paths for pulses: in regard to new food applications, particularly in the bakery sector, and their rising recognition for health [60]. It suggests consolidating research investment on these food applications to sustain their market development and contribute to sustainable legume-based food value-chains in Europe [6].

Secondly, comparatively speaking, certain other topics seem to decelerate although they are strategic for the development of pulses. This is particularly noticeable in the case of works on starches. As starches are a main compound of pulses, whilst proteins attract more attention, both compounds need to be valorised from an economic point of view to ensure sufficient profitability.

Thirdly, this map reveals that some topics reflect few studies such as fibre intake, even though a high consumption of fibre is an important issue for healthy diets. Likewise, the topics of germination belong to a decelerating cluster. As for meat replacement, very few works were revealed. Identifying such decelerating or under-investigated topics could suggest new R&D investment for the pulse sector.

Fourthly, these works are mainly positioned on a scale of ingredients or foodstuffs, less so on a scale of actual meals and not at all on whole diets. Pulses are recommended in sustainable diets as food vectors of protein, fibre, several micronutrients (folates, iron, zinc, magnesium) and other bioactive phytochemicals, such as saponins, phytates and tannins displaying potential beneficial cardio-metabolic and anti-carcinogenic effects. However, those phytochemicals could have antinutritional properties, i.e tannins mentioned as contributors to the higher prevalence of iron deficiency, that may occur in vegan diets.

In that way, revealing the FS&T paths on pulses helps to think about the futures of pulses, to overcome old-fashioned image and low consumption habits by creating new and attractive pulse-based foodstuff [7].

Beyond pulses, which analysis of the FS&T field?

This work also questions the use of scientometric methods to analyse the research paths within the FS&T field. Increasingly developed in recent years, scientometric methods can be employed more systematically [30-31] but this questions the ways : *(i)* to associate experts from FS&T and information scientists; *(ii)* to develop dedicated platforms that can provide scientific field maps through dedicated algorithms, such as those we developed for the present work; *(iii)* to include thesauri sharing common terms of interest that cover the various names of species or compounds studied, and the terms of the scientific field investigated by concatenating synonyms and related words, false friends, frequent typographical errors, etc.; *(iv)* to retrieve large bibliometric datasets (using those thesauri) connected to text-mining algorithms, in order to analyse the semantic structure of a field. The growth rate of scholarly publications calls for efficient and relevant tools to analyse larger and larger corpora of bibliometric data. There has been some progress made on this last point through the development of new platforms, such as the IFSA library or the "EBSCO Food Science Source". The thesaurus used to establish these platforms, however, are still not directly accessible for researchers, so much progress is required to allow downloading large bibliometric dataset as we did, for large-scale analyses. Our work was made possible thanks to a paid subscription to the Expanded API of the WoS [29]. Otherwise IFSA or EBSCO platforms do not allow downloading such large bibliometric datasets. Consequently, at present time, no applications in those bibliometric databases allow detecting scientific trajectories, forefront topics but also under-investigated or decelerating fields. In addition, Dimensions [61] as a new bibliographic platform with a coverage larger than the one of WoS or Scopus particularly for recent years [62], is an interesting source of bibliometric data that could be mined to update our analysis in the future.

Second, our method makes it possible to classify a large dataset of scholarly documents into topics that can be further explored for a specific review of the literature by : *(i)* considering actual "text segments" and not just words isolated from their text context; *(ii)* combining both scientometrics, STS, and FS&T experts through strong interdisciplinary interactions. In this way, this work opens up also new ways for conducting reviews of the literature.

Conclusion

We developed a roadmap for mapping the field of FS&T on pulses which are strategic species for the sustainability of agrifood systems. This work achieved a larger coverage (more than 2,000 scholarly documents analysed over nearly four decades) compared to classic reviews of the literature that generally focuses on one specific topic. Through a lexical and semantic analysis of the abstracts and titles, we analysed the structure of this European research field from 1980 to 2018. This first work opens several perspectives. On the one hand it opens a research agenda on the way to analyse the structuration of the FS&T field beyond the subject of pulses. On the other hand, it opens a research agenda to go further on the analysis on current research topics on pulses over the last decade, compared with this past evolution of pulses. In particular, it will be interesting to evaluate if some research gaps we identified with the recent research topics developed in the decade 2010, for sustaining pulses development in food, are under progress in the decade 2020. In addition, scientometric tools development allows to investigate other dimensions such as understanding the topics on pulses linked to specific scientific communities [63-64], or how some geographic areas are specialized in specific topics [65;37]. All those initiatives will help to define sound Science and Innovation policies for pulses development. The methodological roadmap we established supports the implementation of such analyses.

Acknowledgements

Authors acknowledge Alice Thomson-Thibault for her helpful comments and English editing of the manuscript. First steps of this work was supported by the European project LEGVALUE and final steps by the French project KING funded by the Region Occitanie.

Appendix

Table A. Pulses area cultivated in Europe in 2018 (EU28) from FAOStats

Species	**Hectares cultivated in thousands**
Peas	880.8
Faba beans	379.2
Beans	206.0
Chickpeas	172.2
Lupins	149.8
Lentils	99.4
Vetches	122.1
Other pulses	316.8
Total pulses	**2 327.8**

References

[1] Hall, B. H., Rosenberg, N. (Éds.). (2011). Handbook of the economics of innovation. Vol. 1. North Holland.

[2] Dosi, G., & Nelson, R. R. (2010). Technical Change and Industrial Dynamics as Evolutionary Processes. In Handbook of the Economics of Innovation (Vol. 1, p. 51-127). Elsevier. https://doi.org/10.1016/S0169-7218(10)01003-8

[3] 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. doi: 10.1016/j.gfs.2021.100520.

[4] 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*(10170), 447-492. https://doi.org/10.1016/S0140-6736(18)31788-4

[5] 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

[6] Balázs B., Kelemen E., Centofanti T., Vasconcelos M. W., Iannetta P. P.M., 2021. Integrated policy analysis to identify transformation paths to more sustainable legume-based food and feed value-chains in Europe, Agroecology and Sustainable Food Systems, 45:6, 931-953, DOI: 10.1080/21683565.2021.1884165

[7] Cusworth, G., Garnett, T. and Lorimer, J., 2021. Legume dreams: The contested futures of sustainable plant-based food systems in Europe, Global Environmental Change, 69, p. 102321. doi: 10.1016/j.gloenvcha.2021.102321

[8] 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/legal-content/EN/TXT/?uri=CELEX:52018DC0757

[9] 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*(7), 793. https://doi.org/10.3390/nu9070793

[10] 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*(1), 58-66. https://doi.org/10.1111/nyas.13308

[11] Poux, X., Aubert, P.-M. (2018). *An agroecological Europe in 2050 : Multifunctional agriculture for healthy eating*. *9*, 74.

[12] Magrini, Marie-Benoit, Béfort, N., & Nieddu, M. (2019). Technological Lock-In and Pathways for Crop Diversification in the Bio-Economy. In *Agroecosystem Diversity* (p. 375-388). Elsevier. https://doi.org/10.1016/B978-0-12-811050-8.00024-8

[13] Magrini, Marie-Benoit, 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*, *2*, 64. https://doi.org/10.3389/fsufs.2018.00064

[14] Magrini, Marie-Benoit, 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 agrifood system. *Ecological Economics*, *126*, 152-162. https://doi.org/10.1016/j.ecolecon.2016.03.024

[15] Magrini, Marie-Benoit, Salord, Tristan, Cabanac, Guillaume, 2022 (in press). The unbalanced development among legume species regarding sustainable and healthy agrifood systems in North-America and Europe: focus on food product innovations, Food Security. DOI : 10.1007/s12571-022-01294-9

[16] Jallinoja, P., Niva, M., & Latvala, T. (2016). Future of sustainable eating? Examining the potential for expanding bean eating in a meat-eating culture. *Futures*, *83*, 4-14. https://doi.org/10.1016/j.futures.2016.03.006

[17] Melendrez-Ruiz, J., Buatois, Q., Chambaron, S., Monnery-Patris, S., & Arvisenet, G. (2019). French consumers know the benefits of pulses, but do not choose them : An exploratory study combining indirect and direct approaches. *Appetite*, *141*, 104311. https://doi.org/10.1016/j.appet.2019.06.003

[18] Sonnino, A. (2016). *LEGUMINOSE DA GRANELLA E RICERCA AGRICOLA*. https://www.researchgate.net/publication/315118818_Leguminose_da_granella_e_ricerca_agricola/link/58cbbab392851c31f6570dac/download

[19] Aiking, H., de Boer, J. (2018). The next protein transition. *Trends in Food Science & Technology*, S0924224418301213. https://doi.org/10.1016/j.tifs.2018.07.008

[20] de Boer, J., Aiking, H. (2019). Strategies towards healthy and sustainable protein consumption : A transition framework at the levels of diets, dishes, and dish ingredients. *Food Quality and Preference*, *73*, 171-181. https://doi.org/10.1016/j.foodqual.2018.11.012

[21] Rawal, V. (2019). *The Global Economy of Pulses*. FAO. https://doi.org/10.4060/I7108EN

[22] van der Weele, C., Feindt, P., Jan van der Goot, A., van Mierlo, B., van Boekel, M. (2019). Meat alternatives : An integrative comparison. *Trends in Food Science & Technology*, *88*, 505-512. https://doi.org/10.1016/j.tifs.2019.04.018

[23] Chardigny, J.-M., Walrand, S. (2016). Plant protein for food : Opportunities and bottlenecks. *OCL*, *23*(4), D404. https://doi.org/10.1051/ocl/2016019

[24] Guéguen, J., Walrand, S., & Bourgeois, O. (2016). Les protéines végétales : Contexte et potentiels en alimentation humaine. *Cahiers de Nutrition et de Diététique*, *51*(4), 177-185. https://doi.org/10.1016/j.cnd.2016.02.001

[25] Lam, A. C. Y., Can Karaca, A., Tyler, R. T., & Nickerson, M. T. (2018). Pea protein isolates : Structure, extraction, and functionality. *Food Reviews International*, *34*(2), 126-147. https://doi.org/10.1080/87559129.2016.1242135

[26] Bashi, Z., McCullough, R., Ong, L., & Ramirez, M. (2019). *Alternative proteins : The race for market share is on*. 11.

[27] Rabobank, 2017, Strong growth global pulse production driven by Indian demand, report available through https://www.rabobank.com/en/press/search/2017/20170905-strong-growth-in-global-pulse-production.html, accessed August 2020

[28] Bessada, S. M. F., Barreira, J. C. M., & Oliveira, M. B. P. P. (2019). Pulses and food security : Dietary protein, digestibility, bioactive and functional properties. Trends in Food Science & Technology, 93, 53-68. https://doi.org/10.1016/j.tifs.2019.08.022

[29] 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(23), 6833. https://doi.org/10.3390/su11236833

[30] 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

[31] Todeschini, R. and Baccini, A., 2016. Handbook of bibliometric indicators: Quantitative tools for studying and evaluating research. John Wiley & Sons.

[32] Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang,

D., & Barabási, A.-L. (2018). Science of science. Science, 359(6379), eaao0185. https://doi.org/10.1126/science.aao0185

[33] 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

[34] Noyons, C. M. (2004). Science Maps Within a Science Policy Context : Improving the Utility of Science and Domain Maps Within a Science Policy and Research Management Context. In H. F. Moed, W. Glänzel, & U. Schmoch (Éds.), Handbook of Quantitative Science and Technology Research (p. 237-255). Springer Netherlands. https://doi.org/10.1007/1-4020-2755-9_11

[35] Alfaraz, P. H., & Calviño, A. M. (2004). Bibliometric study on food science and technology : Scientific production in Iberian-American countries (1991-2000). Scientometrics, 61(1), 89-102. https://doi.org/10.1023/B:SCIE.0000037365.53469.91

[36] Seetharam, G., Ravichandra Rao, I. K. (1999). Growth of food science and technology literature : A comparison of CFTRI, India and the world. Scientometrics, 44(1), 59-79. https://doi.org/10.1007/BF02458478

[37] Acosta, M., Coronado, D., Ferrándiz, E., León, M. D., & Moreno, P. J. (2017). The geography of university scientific production in Europe : An exploration in the field of Food Science and Technology. Scientometrics, 112(1), 215-240. https://doi.org/10.1007/s11192-017-2355-5

[38] Borsi, B., Schubert, A. (2011). Agrifood research in Europe : A global perspective. Scientometrics, 86(1), 133-154. https://doi.org/10.1007/s11192-010-0235-3

[39] Guerrero-Bote, V. P., Moya-Anegón, F. (2015). Analysis of Scientific Production in Food Science from 2003 to 2013 : Analysis of scientific production in Food Science, Journal of Food Science, 80(12), R2619-R2626. https://doi.org/10.1111/1750-3841.13108

[40] Guerrero-Bote, V. P., Olmeda-Gómez, C., De-Moya-Anegón, F. (2016). La ciencia de los alimentos georreferenciada. Aproximación bibliométrica a nivel institucional. El Profesional de la Información, 25(1), 25. https://doi.org/10.3145/epi.2016.ene.04

[41] Zhou, P., Zhong, Y., & Yu, M. (2013). A bibliometric investigation on China–UK collaboration in food and agriculture. Scientometrics, 97(2), 267-285. https://doi.org/10.1007/s11192-012-0947-7

[42] Vijay, K. R., & Raghavan, I. (2007). Journal of Food Science and Technology : A bibliometric study. Annals of Library and Information Studies, 54, pp.207-212.

[43] Magrini, M-B., Salord, T., Cabanac, G., Amiot-Carlin, M.-J., Anton, M., Boire, A., Chardigny, J.-M., Larré, C., Micard, V., Nguyen-Thé, C., & Walrand, S. (2020). European Food Science and Technology on pulses : A bibliometric dataset (1980-2018) [Data set]. Portail Data INRAE. https://doi.org/10.15454/JB8LRA

[44] Zohary, D., Hopf, M., & Weiss, E. (2012). Domestication of Plants in the Old World : The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin. OUP Oxford.

[45] Chandra Y. 2018. Mapping the evolution of entrepreneurship as a field of research (1990–2013): A scientometric analysis. PLoS ONE 13(1): e0190228. https://doi.org/10.1371/journal.pone.0190228

[46] Johnson R., Watkinson A., Mabe M. 2018. The STM report. An overview of scientific and scholarly publishing, 5th edition.

[47] Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms. In C. C. Aggarwal, C. Zhai (Éds.), Mining Text Data (p. 77-128). Springer US. https://doi.org/10.1007/978-1-4614-3223-4_4

[48] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining : Classification, Clustering and Extraction Techniques. arXiv:1707.02919 [cs]. http://arxiv.org/abs/1707.02919

[49] Shadrova, A. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices, Journal of Data Mining & Digital Humanities. 7595. https://doi.org/10.46298/jdmdh.7595

[50] Reinert M. 1990. Alceste une méthodologie d'analyse des données textuelles et une application: Aurelia De Gerard De Nerval. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique. 1990;26(1):24-54. doi:10.1177/075910639002600103

[51] Sbalchiero, S., 2018. Topic Detection: a Statistical Model and a Quali-Quantitative Method, Tracing the Life Cycle of Ideas in the Humanities and Social Sciences, Springer International Publishing. https://doi.org/10.1007/978-3-319-97064-6. Chap 10, pp.202-203

[52] Ratinaud, P., Smyrnaios, N., Figeac, J., Cabanac, G., Fraisier, O., Hubert, G., Pitarch, Y., Salord, T. et Thonet, T. (2019). Structuration des discours au sein de Twitter durant l'élection présidentielle française de 2017: Entre agenda politique et représentations sociales, Réseaux, 214-215: 171

[53] Ratinaud, P., Marchand, P. (2012). Application de la méthode ALCESTE aux "gros" corpus et stabilité des "mondes lexicaux" : analyse du "CableGate" avec IRAMUTEQ. In: *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles* (pp. 835-844). Liège, Belgique.

[54] Hoover, R., Hughes, T., Chung, H. J., & Liu, Q. (2010). Composition, molecular structure, properties, and modification of pulse starches : A review. Food Research International, 43(2), 399-413. https://doi.org/10.1016/j.foodres.2009.09.001

[55] Thakur, S., Scanlon, M. G., Tyler, R. T., Milani, A., & Paliwal, J. (2019). Pulse Flour Characteristics from a Wheat Flour Miller’s Perspective : A Comprehensive Review: Pulse flour characteristics…. Comprehensive Reviews in Food Science and Food Safety, 18(3), 775-797. https://doi.org/10.1111/1541-4337.12413

[56] Boukid, F., Zannini, E., Carini, E., & Vittadini, E. (2019). Pulses for bread fortification : A necessity or a choice? Trends in Food Science & Technology, 88, 416-428. https://doi.org/10.1016/j.tifs.2019.04.007

[57] Avilés-Gaxiola, S., Chuck-Hernández, C., & Serna Saldívar, S. O. (2018). Inactivation Methods of Trypsin Inhibitor in Legumes : A Review: Inactivation of trypsin inhibitor…. Journal of Food Science, 83(1), 17-29. https://doi.org/10.1111/1750-3841.13985

[58] Sandberg, A.-S. (2002). Bioavailability of minerals in legumes. British Journal of Nutrition, 88(S3), 281-285. https://doi.org/10.1079/BJN/2002718

[59] Sha, L., Xiong, Y. L. (2020). Plant protein-based alternatives of reconstructed meat : Science, technology, and challenges. Trends in Food Science & Technology, 102, 51-61. https://doi.org/10.1016/j.tifs.2020.05.022

[60] Dahl, W.J. (Ed.), 2019. Health Benefits of Pulses. Springer, Cham, Switzerland.

[61] Herzog, Ch., Hook, D., Konkiel S., 2020. Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies, 1 (1): 387–395. doi: https://doi.org/10.1162/qss_a_00020

[62] Singh, V.K., Singh, P., Karmakar, M. et al. 2021. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics 126, 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

[63] Baccini A., Barabesi L., Khelfaoui M., Gingras Y., 2020, Intellectual and social similarity among scholarly journals: An exploratory comparison of the networks of editors, authors and co-citations. Quantitative Science Studies, 1 (1): 277–289. https://doi.org/10.1162/qss_a_00006

[64] Larivière, V., Gingras, Y., Sugimoto, C.R. and Tsou, A., 2015. Team size matters: Collaboration and scientific impact since 1900. Journal of the Association for Information Science and Technology, 66(7), 1323-1332. https://doi.org/10.1002/asi.23266

[65] Chinchilla-Rodríguez, Z., Sugimoto C. R., Larivière V. 2019. "Follow the leader: On the relationship between leadership and scholarly impact in international collaborations." PloS one 14, no. 6: e0218309. https://doi.org/10.1371/journal.pone.0218309