Peer-Reviewed Literature on Grain Legume Species in the WoS (1980–2018): A Comparative Analysis of Soybean and Pulses

Marie-Benoît Magrini, Guillaume Cabanac, Matteo Lascialfari, Gaël Plumecocq, Marie-Josèphe Amiot, Marc Anton, Gaëlle Arvisenet, Alain Baranger, Laurent Bedoussac, Jean-Michel Chardigny, et al.

To cite this version:

Marie-Benoît Magrini, Guillaume Cabanac, Matteo Lascialfari, Gaël Plumecocq, Marie-Josèphe Amiot, et al.. Peer-Reviewed Literature on Grain Legume Species in the WoS (1980–2018): A Comparative Analysis of Soybean and Pulses. Sustainability, 2019, 11 (23), pp.6833. 10.3390/su11236833 . hal-02416411

HAL Id: hal-02416411
https://hal.science/hal-02416411

Submitted on 26 May 2020

sustainability

MDPI

Article

Peer-Reviewed Literature on Grain Legume Species in the WoS (1980–2018): A Comparative Analysis of Soybean and Pulses

Marie-Benoît Magrini [1,*], Guillaume Cabanac [2], Matteo Lascialfari [1], Gael Plumecocq [1], Marie-Josephe Amiot [3], Marc Anton [4], Gaelle Arvisenet [5], Alain Baranger [6], Laurent Bedoussac [7], Jean-Michel Chardigny [8], Gérard Duc [9], Marie-Hélène Jeuffroy [10], Etienne-Pascal Journet [1,11], Hervé Juin [12], Colette Larré [4], Hugues Leiser [13], Valérie Micard [14], Dominique Millot [9], Marie-Laure Pilet-Nayel [6], Christophe Nguyen-Thé [15], Tristan Salord [1], Anne-Sophie Voisin [9], Stéphane Walrand [16] and Jacques Wery [17]

[1] AGIR, INRA, Université de Toulouse, 31326 Castanet-Tolosan, France; matteo.lascialfari@inra.fr (M.L.); gael.plumecocq@inra.fr (G.P.); etienne-pascal.journet@inra.fr (E.-P.J.); tristan.salord@inra.fr (T.S.)

[2] IRIT, Université de Toulouse, 31062 Toulouse, France; guillaume.cabanac@univ-tlse3.fr

[3] MOISA, CIRAD, CIHEAM-IAAM, INRA, Montpellier SupAgro, Université Montpellier, 34060 Montpellier, France; marie-josephe.amiot-carlin@inra.fr

[4] BIA, INRA, 44000 Nantes, France; marc.anton@inra.fr (M.A.); colette.larre@inra.fr (C.L.)

[5] CSGA, AgroSup Dijon, CNRS, INRA, Université Bourgogne-Franche-Comté, 21000 Dijon, France; gaelle.arvisenet@agrosupdijon.fr

[6] IGEPP, INRA, Agrocampus-Ouest, Université Rennes 1, 35653 Le Rheu, France; alain.baranger@inra.fr (A.B.); marie-laure.pilet-nayel@inra.fr (M.-L.P.-N.)

[7] AGIR, INRA, ENSFEA, Université de Toulouse, 31326 Castanet-Tolosan, France; laurent.bedoussac@inra.fr

[8] DPTI, INRA, 75338 Paris, France; jean-michel.chardigny@inra.fr

[9] Agroecologie, INRA, AgroSup Dijon, Université Bourgogne France-Comté, 21000 Dijon, France; gerard.duc@inra.fr (G.D.); dominique.millot@inra.fr (D.M.); anne-sophie.voisin@inra.fr (A.-S.V.)

[10] Agronomie, INRA, AgroparisTech, Université Paris-Saclay, 78850 Thiverval-Grignon, France; marie-helene.jeuffroy@inra.fr

[11] LIPM, Université de Toulouse, INRA, CNRS, 31326 Castanet-Tolosan, France

[12] EASM, INRA, 17700 Surgères, France; herve.juin@inra.fr

[13] IST, INRA, 84140 Avignon, France; hugues.leiser@inra.fr

[14] IATE, INRA, Montpellier SupAgro, CIRAD, Université Montpellier, 34060 Montpellier, France; valerie.micard@supagro.fr

[15] SQPOV, INRA, 84140 Avignon, France; christophe.nguyen-the@inra.fr

[16] INRA, UNH, Unité de Nutrition Humaine, CRNH Auvergne, Université Clermont Auvergne, 63001 Clermont-Ferrand, France; stephane.walrand@inra.fr

[17] ICARDA, Ismail El-Shaaer Building, Maadi 11728, Egypt; J.Wery@cgiar.org

* Correspondence: marie-benoit.magrini@inra.fr; Tel.: +33-561-285-422

Received: 25 September 2019; Accepted: 22 November 2019; Published: 2 December 2019

Abstract: Grain-legume crops are important for ensuring the sustainability of agrofood systems. Among them, pulse production is subject to strong lock-in compared to soya, the leading worldwide crop. To unlock the situation and foster more grain-legume crop diversity, scientific research is essential for providing new knowledge that may lead to new development. Our study aimed to evaluate whether research activity on grain-legumes is also locked in favor of soya. Considering more than 80 names grouped into 19 main grain-legume species, we built a dataset of 107,823 scholarly publications (articles, book, and book chapters) between 1980 and 2018 retrieved from the Web of Science (Clarivate Analytics) reflecting the research activity on grain-legumes. We delineated 10 scientific themes of interest running the gamut of agrofood research (e.g., genetics, agronomy, and nutrition). We indexed grain-legume species, calculated the percentage of records for each one, and conducted several analyses longitudinally and by country. Globally, we found an unbalanced research output: soya remains the main crop studied, even in the promising field of food sciences advanced by FAO as the "future of pulses". Our results raise questions about how to align research priorities with societal demand for more crop diversity.

Keywords: knowledge dynamics; science and technology studies (STS); research trajectories; scientometrics; bibliometric data; agricultural sciences; food sciences; sustainability

1. Introduction

A major challenge for agriculture is to greatly reduce the use of synthetic inputs and to rely on more ecological processes. To foster sustainability, increasing crop diversity with legumes has been advanced as a major driver of change [1–3]. More legume cultivation will provide considerable ecological services, in addition to their nutritional advantages, as legumes enable a renewable input of nitrogen (N) into agricultural soils through symbiotic N_2 fixation, lowering the fertilizer N requirements and fossil energy use of farming systems, and thus reducing the net release of greenhouse gases into the atmosphere [4]. To obtain these ecosystem services, farming systems must increase crop diversity with more legumes cultivated both in quantity and in number of legume species cultivated. Moreover, crop diversity is linked to the capacity of markets to use a variety of crops for feed and food [5,6]. For soya, huge markets exist, whereas strong innovation is still needed for pulses to break out of the lock-in they face compared to global major crops [7–9]. Even though today legumes are increasingly promoted for human protein intake, enabling a reduction in animal-based consumption, pulses have had difficulties to develop in those markets [7].

Within the same crop family—grain-legumes—pulses and soya have developed very differently since the 1960s. Soya is the main worldwide protein crop grown mainly for feed use (with soya oil increasingly becoming a byproduct). Nowadays, the soya crop amounts to more than 300 million metric tons, while other main grain-legumes such as pulses (pea, lentils, lupins, faba beans, chickpeas, etc.) account for less than 100 million metric tons (Appendix A, Table A1). At global scale, pulse production has gained little ground, whether for food or for feed. Thus, the United Nations launched an International Year of Pulses (IYP) in 2016 to raise awareness about the considerable contributions pulses can make to the sustainability transition of agrofood systems, and to favor their development compared to major crops such as soya [10,11].

To contribute to such new trajectories for grain-legumes, there is a consensus in Science and Technology Studies (STS) that increasing scientific research is essential. Science provides a stock of knowledge for driving new developments and innovation [12–14]. Hence, the main objective of our study is to assess the shares of research activity among the grain-legumes crops in order to evaluate whether research on grain-legumes is itself locked around soya or provides knowledge on a variety of species. Measuring science has been an ambitious challenge for many authors in STS since the seminal works of the 1980s [14]. Bibliometric and scientometric methods are now well-established for analyzing scientific advancement and for orientating research policy [15–20]. The considerable development of scientific platforms [21], combined with algorithmic advances for performing bibliometric analyses, has furthered interest in exploring the dynamics of scientific knowledge [19,22]. Through knowledge and scientific analysis, gaps and opportunities in science can be identified and addressed to meet society's needs for innovation. For sustainability transition studies, bibliometric analysis enables us to better understand the state of knowledge of the current sociotechnical regime that needs to be changed, particularly in sectors with strong path-dependency. Consequently, the number of bibliometric analyses has been growing for a variety of fields: in electronics [23], on scientific parks and the links to open innovation [24], or in pharmaceuticals [17]. However, few have been done in agriculture (e.g., on agroecology [25]), even though there are key sustainability issues challenging research.

To evaluate the respective shares of grain-legume species in the sciences, we used the Web of Science (WoS) collection to retrieve the scholarly publications reflecting research activity recognized by peers. As advanced by STS, and more specifically by scientometrics studies [19], these research papers remain the highest quality variable to reflect research activity on a given topic. Even though GS (Google Scholar) is growing and presented as alternative, the metadata offered by GS are still very limited, reducing the practical suitability of this source for large-scale citation analyses. In addition, GS includes around 20–50% of other type of research documents (depending on the scientific fields), such as PhD dissertations, scientific reports, and non-peer-reviewed papers that create bias in comparing research advancement from more qualitative types of research documents [26]. Moreover, language harmonization rules are required to collect data on GS. Therefore, although perhaps not a perfect resource to reflect all the scientific knowledge stock, we chose a traditional bibliometric platform, as the WoS as better for providing more a relevant overview of peer-reviewed research activity (i.e., the scholarly documents such as articles in peer-reviewed journals, books and book chapters), 97% of which were written in English on the WoS. An alternative bibliometric platform is Scopus, but we chose WoS thanks to our institution access; and Scopus use will generate additional computing methods to extract large collection as we did on the WoS, without adding so much other records as WoS and Scopus present high overlap level [26].

We chose to focus on temperate climate species; that is, the grain-legume species grown in most Western countries; but many of them are also important grain-legumes for semi-arid or tropical areas. In analyzing the scholarly publications on grain-legumes, we wanted to assess the relative occurrences of several grain-legume species and which countries had been the most involved in this research. We performed this analysis for research over the last four decades (between 1980 and 2018), by considering ten themes (i.e., scientific fields) of interest in agrofood system research and identified by experts: Genetics, Agronomy, Ecophysiology, Biotic-stress, Feeding, Processing, Nutrition, Allergy, Acceptability, and Socioeconomics. No review of the literature to date has managed to tackle so many themes of interest together for both agricultural and food sciences, whatever the crop species considered.

This original and ambitious literature review involved 26 scientific experts on legumes, coupled with database and scientometrics experts to create relevant search queries on the WoS and appropriate software to process and analyze the resulting bibliographic records. Our findings are based on a core corpus totaling 107,823 scientific publications (i.e., records) retrieved by thematic search queries addressing the *title*, *abstract*, and *authors' keywords*. Since soya is a major crop used for oil unlike most other grain-legumes, we excluded records referring to the subject of "soya oil", in order to have more relevant comparisons in the corpus created. Our results show that soya is mentioned in 43% of the records, groundnut in nearly 10% and all other pulses combined in 47%. The analyses revealed a strong imbalance within grain-legume species research, with soya dominant over all other grain-legume species; and if "soya oil" had been included in keyword searches, this percentage would be even greater. This trend has grown even stronger in recent years. We also observed that the breakdown of themes researched were not the same for soya and pulses. Processing and Nutrition were much more common themes of research for soya than for pulses. For pulses, research mainly focused on "upstream" themes linked to Genetics or Ecophysiology. This imbalance questions the capacity of research to develop knowledge that would enable more food outlets for pulses, as expected by the United Nations during the IYP.

2. Materials and Methods

2.1. Overview of the Methodology Adopted

This study was based on a bibliometric dataset retrieved from the Web of Science (WoS) a product of Clarivate Analytics. As mentioned in Section 1, Scopus is the other prominent bibliographic database, which presents a high overlap with the WoS: see [26] for a comparison of these databases stressing their limitations and advantages to conduct large-scale literature analyses. The WoS is one of the most used bibliometric resources in the world. It provides access to article records from more than 30,000 journals and books in various fields of science. The WoS "Core Collection" includes about 70 million records [27].

For the present study, we worked with several scientific experts to identify domain-specific keywords and retrieve the associated scientific publications (records). First, we identified keywords covering most of the cultivated grain-legume species in Western countries (Species query). Second, we determined 10 domain-specific themes covering the main research issues on grain-legumes (e.g., Allergy). Then, search queries on the WoS were designed to delineate the species and each theme (Table 1), leading to 10 thematic corpora on grain-legumes. Finally, these 10 corpora were merged into a single corpus called Fusion. All search queries and the associated publication records are available as open data from the https://data.inra.fr repository (Appendix A, Table A2).

In this article, a "corpus" is a set of records retrieved from the WoS and a "record" refers to the metadata of a scholarly document (e.g., the type of publication, such as "article", "book chapter"; the authors; etc.). Designing these queries was an iterative process requiring a combination of skills:

- Theme experts

  We identified leading scientists in several fields (Table 1) who helped to delineate search queries for 10 key themes and to check the validity of the records retrieved. (Delineation of scientific fields or subject areas is a question under much discussion in scientometrics. Bibliometric databases developed their own classification reflecting the scope of journals. As such classification were not directly suitable to break down agricultural and food research activity in several fields on which to build our search queries, we relied on experts' judgement to define 10 themes of main interest covering the research on grain-legumes. We used the term of "theme" as a synonym of "subject area". See [20] to go further on this question of delineation of scientific fields).
- Scientometricians

  We collaborated with scientists who specialize in the quantitative study of science and database management systems to create an online platform giving access to the corpora collected. This platform was used to then incrementally refine search queries to build the corpora. Since the number of records to collect from the WoS (100 k) exceeded the amount of records one can extract from the web interface (limited to 5 k), we resorted to a Web of Science Data Integration feature. Data collection was then performed with an in-house program using the Web of Science API Expanded.

Table 1. Thematic corpora investigated by experts.

Theme and Underlying Corpus Name	Description of the Theme	Number of Scientific Experts Involved
Species	Names used to designate the various main grain-legume species and varieties cultivated in temperate climates	2
Genetics	Varieties, genes, breeding methods and objectives	2
Agronomy	Ways to grow legume crops and provided services	2
Ecophysiology	Plant physiology in relation to its abiotic environment	2
BioticStress	Weeds, diseases, and pests' life traits and control in crops	2
Feeding	Feeding practices, animal nutrition	2
Processing	Transformation and main types of food products excluding non-food uses	4
Nutrition	Nutrition subjects for humans including health	4
Allergy	Concerns on allergy linked to the use of legumes in food	2
Acceptability	Sensorial and organoleptic analysis for consumer acceptance	2
Socioeconomics	Any subject of interest using socio-economic approaches	2

Regarding the 10 domain-specific themes selected, we adopted a broad-spectrum approach covering the end-to-end workflow of grain-legume research from production to consumers. A growing concern in Western countries is to increase legume consumption [28]; therefore, we made sure to cover allergy and consumer acceptance subjects. We also designed a query capturing the main socioeconomic research theme on grain-legumes (e.g., funding of work on breeding activities, farmers' production choices, feed practices and business, consumer behaviors, and market functioning, policies). This query led to the corpus called Socioeconomics, a research theme complementary to the others on life sciences and engineering.

Performing the statistics at the meta-level required merging the 10 thematic corpora on grain-legumes into a single corpus called Fusion. We merged all records from the underlying corpora, removing duplicates that we identified thanks to the unique identifier provided for each record in the WoS (UT code). As a result, a record in Fusion appeared only once, even if it appeared in multiple underlying corpora.

2.2. Designing Search Queries on the WoS: Main Principles

This section introduces the main principles we followed to design the search queries (Table A2) (for the syntax rules of search queries, see the *Web of Science Core Collection Help*: http://images.webofknowledge.com/WOKRS5251R3/help/WOS/hp_search.html).

2.2.1. Documents Type

The search was restricted to the main types of scientific literature documents, namely: article, book, book chapter, and review. This translated into the query: DT = ("Article" OR "Book" OR "Book Chapter" OR "Review").

2.2.2. Time Range

We selected the period 1980–2018 (PY = "1980–2018") to observe long-term dynamics. This timeframe encompasses document records of variable completeness: abstracts were available only starting in 1990. We expected an increase in the volume per year to raise around 1990 because search queries would match more records in both the titles and abstracts.

2.2.3. Indexing as a Post-Processing Step to Cleanse the WoS Results

WoS queries were issued with the standard *TS* operator, meaning *Topic Search*. *TS* results collect records of the database that match the user's query on the following criteria: title, abstract, author keywords, and *KeyWords Plus*. *KeyWords Plus* are additional terms automatically generated by the WoS and attached to the records. Terms appearing more than once in the titles of the cited references in a record (i.e., in its bibliography section) are called *KeyWords Plus* [29]. For instance, a search like TS = "protein AND pea" matches documents whose title, abstract or author's keywords contain the term "protein" but not "pea", because the term "pea" occurred only in the bibliography of the document retrieved ("pea" being a *KeyWords Plus* here) (see, for instance, query UT = "WOS:000226807600008" AND TS = ("protein" AND "pea") yielding one journal paper on rice mutants, whose abstract contains "protein" but not "pea"). This example stresses that the records retrieved from the WoS might contain records that were not of direct interest. In our point of view, a document of direct interest (to build a core corpus) must contain both terms in the main contents provided only by the authors: title, abstract, and author's keywords.

To circumvent the biases introduced by *KeyWords Plus*, we designed an indexing algorithm that cleaned the WoS results by filtering out non-direct interest documents. Each remaining document had to match our criteria: only author-specified contents should be matched with the query. We applied this indexing algorithm to the 10 thematic queries.

2.2.4. Interactive Browsing of the Bibliographic Corpora

To ease team work with the thematic experts, we designed an online interactive bibliometric platform called SCIM, enabling them to explore the resulting records. The thematic corpora are shown with a clear distinction between direct interest vs. non-direct interest documents. Skimming through the latter allowed experts to check that those eliminated records were clearly out of scope (i.e., validation of the aforementioned indexing step). In addition, using descriptive statistics, experts were instructed to check the relevance of the records included. They checked the most frequent terms, journals, and Web of Science categories (wcs in the WoS parlance (http://images.webofknowledge.com/WOKRS5251R3/help/WOS/hp_subject_category_terms_tasca.html)) associated with the records of the direct interest documents.

2.2.5. Iterative Design and Validation of the Search Queries

Each thematic search query was stabilized with an iterative validation process involving the experts at each iteration. One iteration involves the following tasks applied to each thematic corpus:

- Checking of a random sample from the thematic corpus.

  Each expert was asked to examine the records from a random selection of the 300 documents published during the last three years and present in the thematic corpus he/she was responsible for. Documents considered irrelevant were analyzed to deduce the changes that needed to be made on the search query in order not to retrieve these in the next iteration. Conversely, experts were also asked to identify those aspects of the theme that were not caught by the query. In particular, experts expected the leading authors or topics to appear (the three last years corresponding to the current state of the art); they used this information to adjust search queries when necessary. Overall, the search operator *t1* NEAR/10 *t2* (i.e., term *t1* must be within 10 words away from term *t2*) proved the most efficient for identifying relevant thematic corpora. In our case, *t1* were names of the species studied while *t2* were terms related to the considered theme. This first task iterated until the percentage of irrelevant documents was less than 20% of the random sample.
- Checking of the entire thematic corpus.

  This second task relied on descriptive statistics. For each thematic corpus, experts were instructed to assess the relevance of the most frequent:

  ○ terms in the title, abstract, authors' keywords of the records; and
  ○ *WoS categories* (wcs) reflecting the scope of the journal or the book that the WoS attributes to each record of the bibliographic database.

Irrelevant terms or wcs with high frequencies relative to the thematic corpora were identified and used to adapt, once again, the search query. In particular, this led the experts to specify excluding conditions (see below).

During this phase, the experts of closely related scientific themes (for instance, between ECOPHYSIOLOGY and AGRONOMY, or between NUTRITION and PROCESSING) collaborated to better delineate each theme. Several meetings with all experts led to adaptations of the queries between themes. This ensured a reduced asymmetry of knowledge between the scientific experts for each theme, which helped to better delineate the search queries.

2.2.6. Excluding Conditions

Queries feature excluding conditions on some keywords or wcs, for specific themes or for all of them.

- For instance, in the PROCESSING query, the "germination" keyword is ambiguous, as it relates to either a food subject or an agronomic subject (as regards the germination step of seeds in the soil concerning more the ECOPHYSIOLOGY corpus). As "germination" was a keyword that we need to keep for the PROCESSING query, we excluded the wcs of the PROCESSING corpus not related to "Nutrition Dietetics" and "Food Science Technology".
- For all queries:

- The terms "coffee" and "cacao" where excluded because they also appear in the underlying paper under the generic term "bean".
- The phrase "soya oil" was excluded due to a twofold rationale. First, it is over-represented in the literature on soya. Second, comparing soya and pulses requires selecting common features of legume interests, such as the increasing interest in plant-based protein development for food instead for oil [28].
- The phrase "biodiesel" and "biofuel" as products linked to oil fraction was excluded. In general, non-food uses are beyond the scope of this study.

Finally, all of these issues meant that conducting a bibliometric study requires careful, in-depth, and iterative expert coordination for performing the many checks and refining the search strategy. Compared to the delineation strategy established for this Fusion corpus (Figure 1), other delineation strategies were also tested and are reported in Appendix B. Those alternative delineation strategies resulted in less relevant corpora than with the delineation strategy of Figure 1 (see Table A5 for instance).

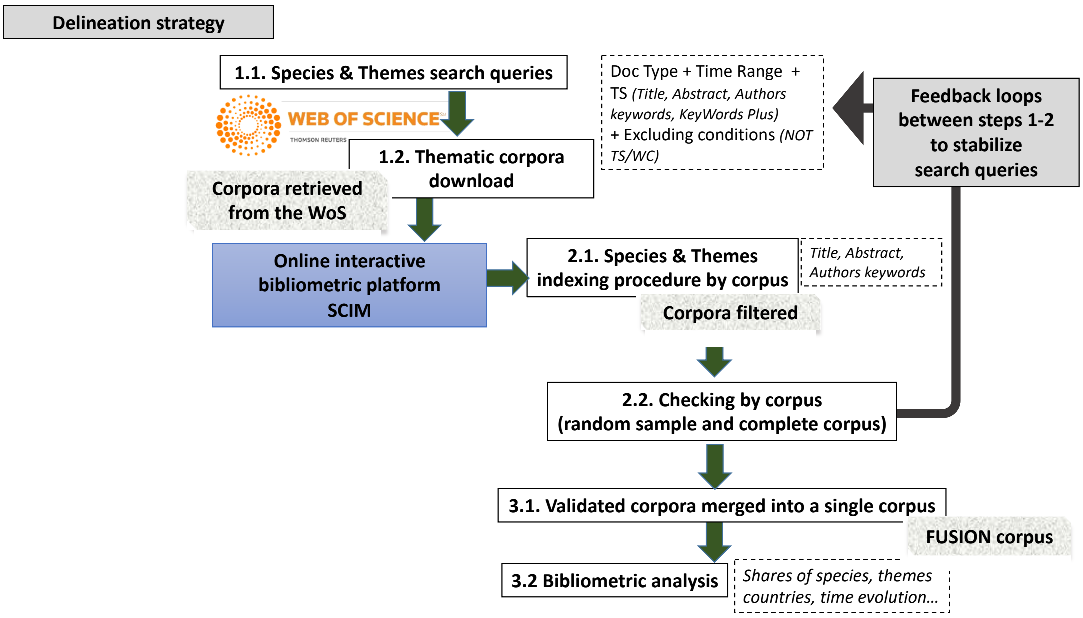

Delineation strategy
1.1. Species & Themes search queries
WEB OF SCIENCE
THOMSON REUTERS
Doc Type + Time Range +
TS (Title, Abstract, Authors keywords, KeyWords Plus)
+ Excluding conditions (NOT TS/WC)
Feedback loops between steps 1-2 to stabilize search queries
1.2. Thematic corpora download
Corpora retrieved from the WoS
Online interactive bibliometric platform SCIM
2.1. Species & Themes indexing procedure by corpus
Title, Abstract, Authors keywords
Corpora filtered
2.2. Checking by corpus (random sample and complete corpus)
3.1. Validated corpora merged into a single corpus
FUSION corpus
3.2 Bibliometric analysis
Shares of species, themes countries, time evolution…

Figure 1. Main steps to build the bibliometric dataset: the delineation strategy.

This figure summarizes the main steps followed to build the thematic corpora. First, queries were submitted to the Web of Science to collect bibliographic records for each theme (Steps 1.1 and 1.2). Second, the indexing phase eliminated the records retrieved because of *KeyWords Plus* only (Step 2.1). Then, for each theme, surviving records were checked by experts: first, on a random sample, and, second, for the complete corpus (Step 2.2). Third, the experts relied on descriptive statistics to identify the remaining irrelevant documents; they modified the query accordingly and re-ran the whole process. In this way, queries were progressively refined. Finally, the 10 thematic corpora validated by experts were merged to form a single corpus of unique records called Fusion (Step 3.1), on which the bibliometric analysis relied (Step 3.2).

2.3. *Focus on the Design of the Species WoS Query*

The 10 thematic search queries were combined with a query targeting the names of grain-legume species. Hence, delineating the various terms referring to grain-legumes was a crucial part of our bibliometric study. This was a challenging task as different names are used in different scientific fields. Either the scientific name (generally the Latin name) or various, country-specific, common names were used. The expertise of two senior researchers internationally recognized on those crops, combined with the consulting of websites dedicated to data collection on plants (Table 2) and books on plant taxonomy (e.g., [30]) enabled us to list more than 80 various names that we grouped into 19 main grain-legumes species (Table 3). Only main pulses, soya, and some species from the genera *Lathyrus* and *Vicia* cultivated in temperate climates (mainly of Western countries and the Mediterranean basin) were considered here. All these species belong to the family of grain-legumes. Among them, according to the United Nations, the grain-legumes not used for oil extraction are commonly called pulses, excluding soya and groundnut for their dual richness in oil and protein. Appendix C gives a brief history of grain-legumes and reveals that while pulses were common crops for centuries, interest in them has dramatically decreased in recent decades compared to soya.

Table 2. Main sources consulted to check the various terms used for grain-legumes species.

Name	Website
Tela Botanica, the French-speaking network of botanists, presenting Latin name of species	https://www.tela-botanica.org/
Feedipedia, describing all the resources used for feed in the world, and among them legumes	https://www.feedipedia.org/
Atlas managed by the CGIAR—Research Program on Dryland Cereals and Legumes Agri-Food Systems (DCL)	http://www.eatlasdcl.cgiar.org
A personal website created by a renowned retired botanist.	https://www.cropsreview.com/grain-legumes.html
Inventory of food resources and constituents	http://foodb.ca

Table 3 presents all the species terms included in the Species search query, classified according to a single species identifier and a generic one. For instance, we gathered under the identifier "soya" all occurrences of various terms referring to soya by using wildcards such as in the following list: glycine max, soja, soya$, soy$, sojabean$, soybean$, and soyabean$. These species identifiers served for the indexing step (explained in Step 2.3) and the exploratory statistics.

As for the thematic search queries, wildcards were used to catch various forms of a term: for instance, being written in plural or singular or with varying country-dependent orthotypography (e.g., soya vs. soja and faba vs. fava bean). The asterisk (*) represents any group of characters (including no character), the dollar sign ($) represents zero or one character. The phrases (expressions composed of several terms), such as "chick pea", were surrounded with quotation marks in the search queries applied on the WoS.

Some generic terms (such as legumes and leguminous) were also considered, but they were only added for the Socioeconomics search query. For the other themes, using these generic terms retrieved too many irrelevant records: that is, records mentioning legumes without dealing specifically with grain-legumes. One explanation for this relates to how scientists phrase their papers: some (especially life sciences) usually work on a specific legume, while others (especially social sciences) tend to consider legumes in a more comprehensive approach.

Table 3. Species identifier and species expressions used in the species search query.

Species Identifier (Genus or Common Name)	All Species or Common Name Terms Included in the Search Query
Adzuki	phaseolus angularis, vigna angularis, red mung$, red bean$, 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$
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$, black gram, blackgram$
Groundnut	arachis hypogea, arachis hypogaea, groundnut$, peanut$
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$
Soya	glycine max, soja, soya$, soy$, sojabean$, soybean$, soyabean$
Vicia	vetch$, vetche$, vicia sativa, vicia villosa, vicia ervilia, ervil$, vicia narbonensis, narbon bean$
Winged bean	psophocarpus tetragonolobus, winged bean$, asparagus pea$, goabean$, goa bean$
Generic	leguminous, *legume, *legumes, pulse, pulses

Note: $ (respectively, *) is a wildcard replacing zero or one character (respectively, no character at all or a group of characters). For instance, "*legumes" matches "grain-legumes" or "dried-legumes".

3. Results and Discussion

Bibliometric analysis allowed us to mine this dataset and infer new knowledge on grain-legume research at the global scale. The main purpose of this paper is to shed light on the share of species in these corpora, reflecting the research activity on grain-legumes. Thus, we calculated the percentage of records on each species within the corpora, longitudinally and at different levels of analysis, according to species, themes, and countries, especially by comparing soya and pulse frequencies. Then, we discuss the implication of these results as regards avenues for future research and policies regarding the question of crop diversity.

3.1. Proportions of Grain-Legume Species in the Scientific Literature

For this subsection, we consider five groups of grain-legume species when reporting our results:

- G1 is for Soya.
- Pulses divided into three groups:
- G2 groups "PFL" including Pea, Fababean, and Lupin species. This group is the European classification of the main protein-rich crops among pulses.
- G3 groups "Other pulses" for the remaining pulses but excluding "Lathyrus and Vicia"
- G4 groups Lathyrus and Vicia species together as the number of records related to those species are very low and currently the least used.
- G5 is for Groundnut (not considered as a pulse because of its oil richness).

Table 4 present the shares of the five groups of grain-legume species captured in the Fusion corpus: G1, G2, and G3 appear in Table 4a, while G4 and G5 appear in Table 4b.

We also distinguished the counts of species co-occurring with generic terms. Indeed, in the indexing procedure (see Section 2), we indexed the records with generic names of species such as "legumes" or "pulses", in order to appreciate how researchers broaden their studies by referring also to the family group of legumes. Therefore, some figures in this subsection present specific counts of the co-occurrence of a generic term with a specific species' name in records. In addition, we report the average annual growth rate of records in the Fusion corpus.

Table 4. (a) Number and percentage of records in the Fusion corpus related exclusively to soya or pulses, broken down by time periods. (b) Number and percentage of records in the Fusion corpus related exclusively to groundnut, Lathyrus or Vicia, broken down by time periods.

(a)

PERIOD (PER.)	TOTAL PER.	1st PER.	1980s	1990s	2nd PER.	2000s	2010s	3 Last Years
Family species index terms of records:	1980–2018	1980–1999	1980–1989	1990–1999	2000–2018	2000–2009	2010–2018	2016–2018
Soya (alone **)	45,615	13,565	4824	8741	32,050	13,141	18,909	7076
Soya and a generic term *	1440	286	7	279	1154	429	725	277
Other Pulses than PFL (alone)	20,780	7394	2719	4675	13,386	5549	7837	2819
Ibidem and a generic term	3175	569	38	531	2606	914	1692	745
PFL (alone)	16,760	7564	2743	4821	9196	4279	4917	1708
PFL and a generic term	2567	546	15	531	2021	789	1232	480
Subtotal (Soya, Pulses)	90,337	29,924	10,346	19,578	60,413	25,101	35,312	13,105
% Soya in subtotal	52%	46%	47%	46%	55%	54%	56%	56%
% Subtotal within the per.	84%	86%	90%	84%	83%	84%	82%	82%
Total records for the period	107,823	34,652	11,454	23,198	73,171	30,017	43,154	16,034
% per. in the Fusion corpus	100%	32%	11%	21%	68%	28%	40%	15%
Annual average growth rate of the records	13%	21%	32%	12%	6%	6%	6%	6%

(b)

PERIOD (PER.)	TOTAL PER.	1st PER.	1980s	1990s	2nd PER.	2000s	2010s	3 Last Years
Family species index terms of records:	1980–2018	1980–1999	1980–1989	1990–1999	2000–2018	2000–2009	2010–2018	2016–2018
Groundnut (alone)	9305	2627	841	1786	6678	2445	4233	1595
% Groundnut within the per.	*9%*	*8%*	*7%*	*8%*	*9%*	*8%*	*10%*	*10%*
Lathyrus or Vicia (alone)	1097	277	67	210	820	330	490	173
% Lathyrus or Vicia within the per.	1%	1%	1%	1%	1%	1%	1%	1%

(a) Note: Each row in the table is exclusive of that group. PFL: pea, faba bean, or lupin. * For instance, this row shows the number of records containing only a generic term (a generic term such as legumes) AND a term linked to soya in the *Title, Abstract or Authors' keywords*. ** The various combinations of species index (associating pulses, soya or groundnut in records) are not included here. For instance, over the entire period, we counted 838 records co-indexed with groundnut and soya; 298 records co-indexed with Lathyrus–Vicia and PFL, etc. These various co-indexations records over the entire period represent 6% of the Fusion corpus. (b) Note: Each row in the table is exclusive of that group.

3.1.1. The Number of Grain-Legume Publications Grew at the Same Rate as All Records in the WoS Core Collection

First, concerning changes in the corpus size over time, we observed that, even though today there is greater awareness of legumes, the growth of scientific publications on grain-legumes is similar to that within the whole WoS Core Collection; the annual growth rate in scholarly peer-reviewed English-language journals of the WoS increased 5–6% in recent decades [27] (p. 25). The specific higher rate in the years 1980 and 1990 is due to the increase in research activity and publications observed in the entire WoS Core collection (due also to the WoS index rule changes since 1990 including both abstracts and keywords), and not to a special interest in legumes. Therefore, these figures firstly show that there has been no particular increase in legumes research, even after the United Nations' communication about the benefits of pulses for sustainability in the 1980s [31] and more recently in the 2010s (IYP in 2016). Nevertheless, researchers are aware that most publications observed in a given year are the results of research conducted over the three or more previous years. Therefore, measuring any impact from IYP 2016 on the number of publications can take at least 10 years, given that, before an increase in research activity, public and private decisions must be taken to increase funding for that research.

Overall, as in the entire Core Collection, the second period (post-2000s) represents nearly two-thirds of the corpus, with a net increase in research activity (as in the WoS Core collection) since 2010: the 2010–2018 period accounts for 40% of the records in the Fusion corpus. This point is important to stress: as there has been a rapid and strong increase in scientific knowledge, it is crucial to adopt tools to analyze the ways that new knowledge is created. We need to assess the risk of over-investigating some themes or species and under-investigating others.

3.1.2. Generic Terms Referring to Legume or Pulse Family Are More Used with Pulse Species than with Soya Species

Second, concerning the frequency of generic term co-occurrence with species indexes, we observed different tendencies for soya and pulses. Table 4a shows that pulse species were more frequently co-indexed with a generic term than was soya. For instance, for the most recent period (2010–2018), 19% of pulse records are co-indexed with a generic term compared to 4% for soya records. Soya is a dominant crop having developed its own identity, making it distinct from other legumes. In other words, it seems that the identity of belonging to a larger family such as legumes is stronger for pulses species than for soya. It reflects also the fact that research studies focus on one species and rarely relate to the broader context of legumes or pulses.

In addition, the results show a low rate of co-indexation between species (6%) (Table 5 and Figure 2). This means that the links between species are rarely stressed by the authors. In particular, only 2% of records were co-indexed both with soya and a pulse species. Thus, researchers rarely considered two (or more) species when investigating an issue, or, at least, the extension or impacts of the results for other species of the same family were rarely mentioned. This finding fundamentally questions the diffusion of knowledge between species and calls for future research using semantic networks to analyze which concepts (i.e., knowledge) are used when establishing connections between species. Moreover, here we considered only the titles, abstracts and keywords; it is possible that the main text of the article mentioned such impacts for species other than the main one under investigation. However, titles and abstracts express the core message of an article, and considering application of results for other species does not seem to be part of this core message for most research work.

3.1.3. Soya Strongly Dominates within Grain-Legume Publications

Third, these frequencies show the predominance of soya among grain-legumes (Figure 2). It is important to note that this soya predominance is underestimated, as all records referring to "soya oil" were eliminated. When such a similar exclusion is performed on groundnut (i.e., excluding "oil" theme), its ranking is lower: among 11, 612 records indexed with "groundnut", 50% refer to oil thematic. Within the subtotal formed by soya and pulse groups (the two main species families of current interest in developed countries, notably for increasing plant-based protein), soya accounts for more than half of all records. This tendency has grown even stronger in recent years, with soya reaching 56% of the records.

Table 5. Number and percentage of the various grain-legume species quoted in Fusion corpus.

Species Index	Number of Records	Share of Species Index
Soya	51,395	42.6%
Pea	14,175	11.7%
Groundnut	11,612	9.6%
Bean	8976	7.4%
Cowpea	5954	4.9%
Chickpea	5373	4.5%
Faba bean	4640	3.8%
Mungbean	3785	3.1%
Lupin	3420	2.8%
Lentil	2724	2.3%
Pigeon Pea	2302	1.9%
Vicia	2036	1.7%
Gram bean	1466	1.2%
Adzuki	771	0.6%
Fenugreek	762	0.6%
Lathyrus	491	0.4%
Lablab	390	0.3%
Winged bean	232	0.2%
Bambara Bean	173	0.1%
Total species quotes	120,677	100%
Nb of records indexed with only one species (%)	100,739 (94%)	
Nb of records co-indexed with several species (%)	7084 (6%)	
Nb of records in FUSION corpus	107,823	

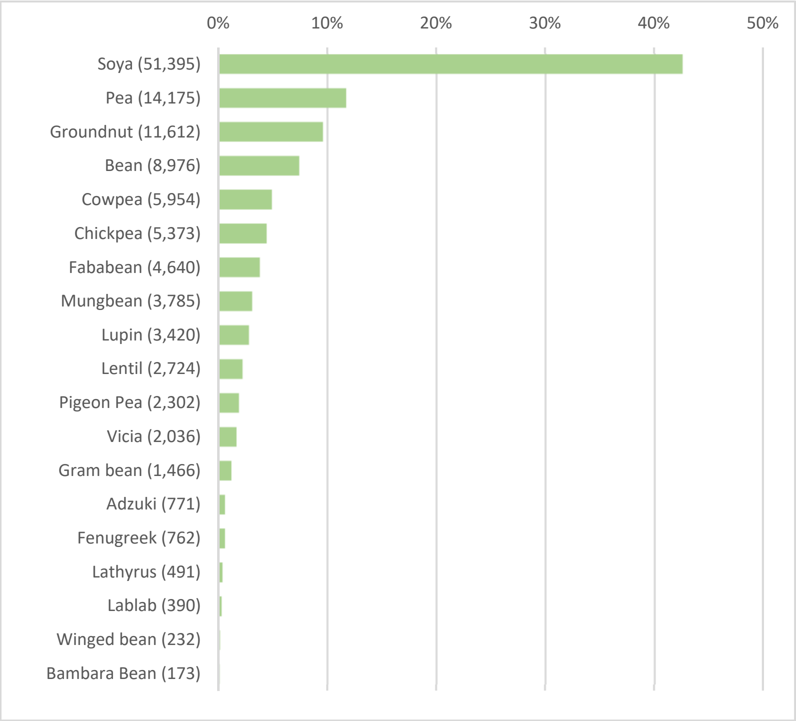


Figure 2. Species distribution quotes in Fusion corpus.

Fourth, when looking for the distribution of all the grain-legume species under study (Table 5), the predominance of soya is more accurate as the second main species, pea, accounts for nearly 12% of the mentions, compared to nearly 43% for soya. Among pulses, the five most mentioned species were, respectively: pea, bean, cowpea, chickpea, and faba bean.

As explained in Appendix B, we created alternative bibliometric corpora (Species1, Species2, Species3). Although in these alternative corpora the themes are less well delineated, it was worthwhile to check whether the species frequencies were similar, since, with the Species3 and Fusion corpora, nearly one-third of the records were not common to both. Finally, we observed the same statistics for these other corpora and the percentage of species was similar (Table A6). Overall, soya accounts for more than half of the scientific publications within the soya and pulse records, with the share of soya increasing in recent years. We also observed in these alternative corpora that soya and pulses represented around 90% of the records on grain-legumes.

3.1.4. Changes in the Mentions of Species over Time

Figures 3 and 4 present the longitudinal evolution of the ten most mentioned species. Soya is the species with the highest number of publications: 45,615 records co-indexed with soya (Table 4a). The increase in records indexed with groundnut, chickpea, lentil, or mungbean was greater in recent years compared to other pulse species.

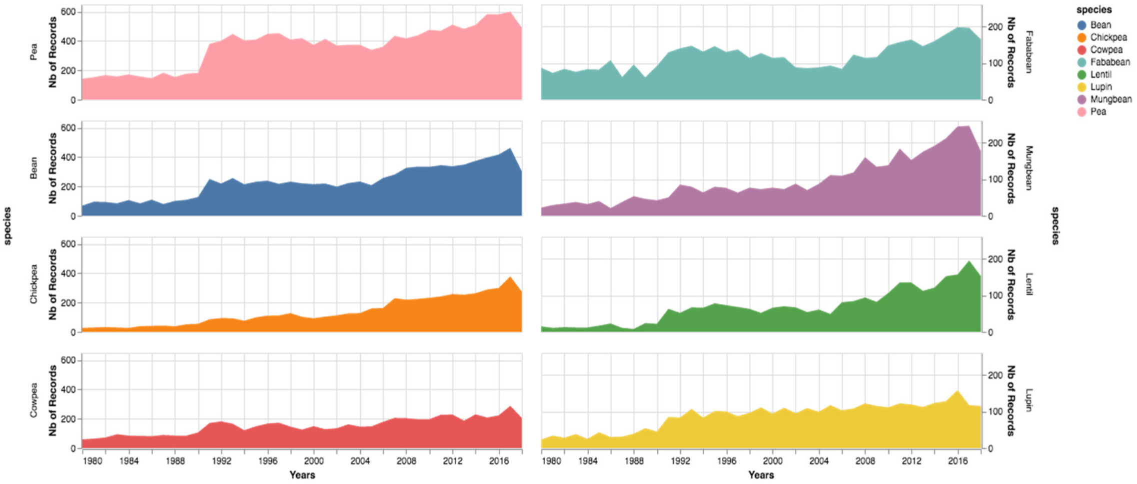


Figure 3. Longitudinal evolution of main pulses species quoted in Fusion corpus.

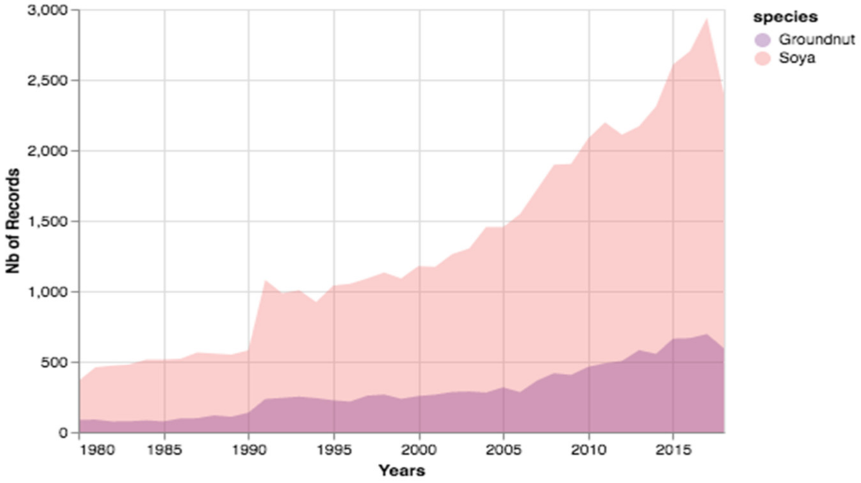


Figure 4. Longitudinal evolution of Soya and Groundnut species quoted in the Fusion corpus.

3.2. Percentage of Grain-Legume Species in Literature across Countries

The metadata of most records retrieved from the WoS identified the countries of the authors. We studied the share of soya and pulses research by country. Figures 5 and 6 and Table 6 report a proportional count for international collaboration records (that is associating several countries):

each country accounts for 1/*n* where *n* is the number of countries associated for the record. As most international collaborations involved few countries, those publishing the most on grain-legumes, this rule avoids overcounting them as when one uses full counting (i.e., 1 point to each country). We observed that with or without proportional counting, the ranking of countries did not change (Figure 5a,b). We also created a specific geographical index of the current 28 European Union countries (whatever the period considered), while computing the count for individual European countries. For the following data, we considered only the records indexed either with soya or with pulses and for which authors' countries were identified (around 15,000 records in the Fusion corpus did not have this information in the WoS; these are mostly papers with multiple authors and only one reprint address, see UT = WOS:A1997XJ61100010, for instance).

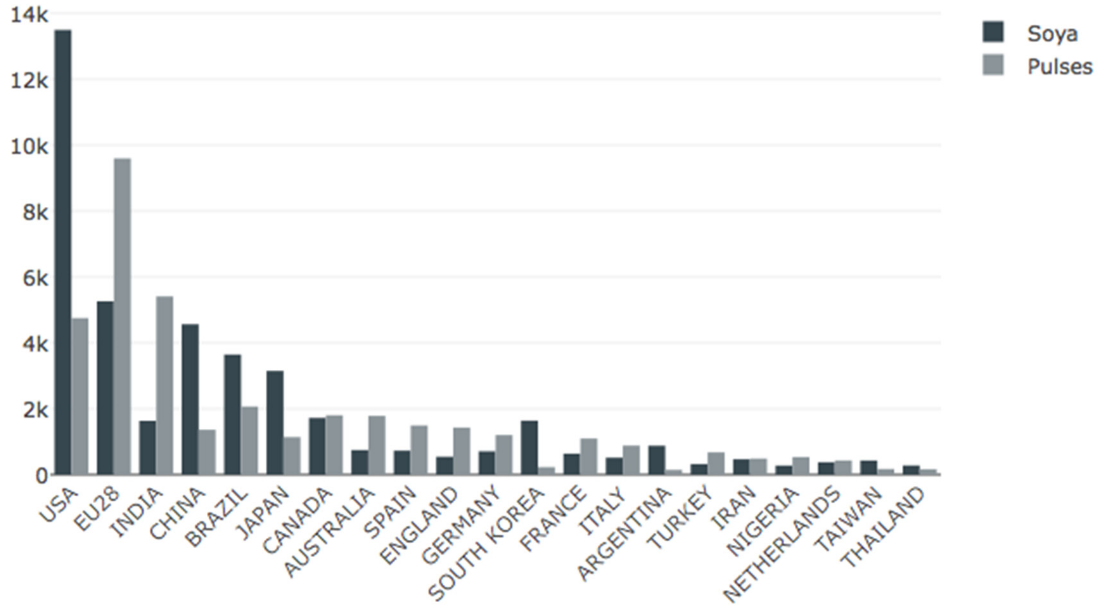
3.2.1. The Ranking of Countries Was Rather Stable over Time, with China, Brazil, and India Rising

First, Figures 5 and 6 show that the ranking of countries was different for soya and pulses, but quite stable over time. Considering both soya and pulses, the two first publishing geographical areas are the USA and the EU28, considering either the whole period or the current decade. The USA and the EU28 account for more than half of the records in the previous period, but less than one-third in the current decade (Table 6). More recently, China has been rising and currently represents 13% of the records in the current decade, followed by India, Brazil, Japan, Canada, and Australia. These seven countries and the EU28 account for two-thirds of the records on soya and pulses over the current decade (versus 84% in the previous period). This reveals a progression of other countries in legume research, such as South Korea and Argentina.

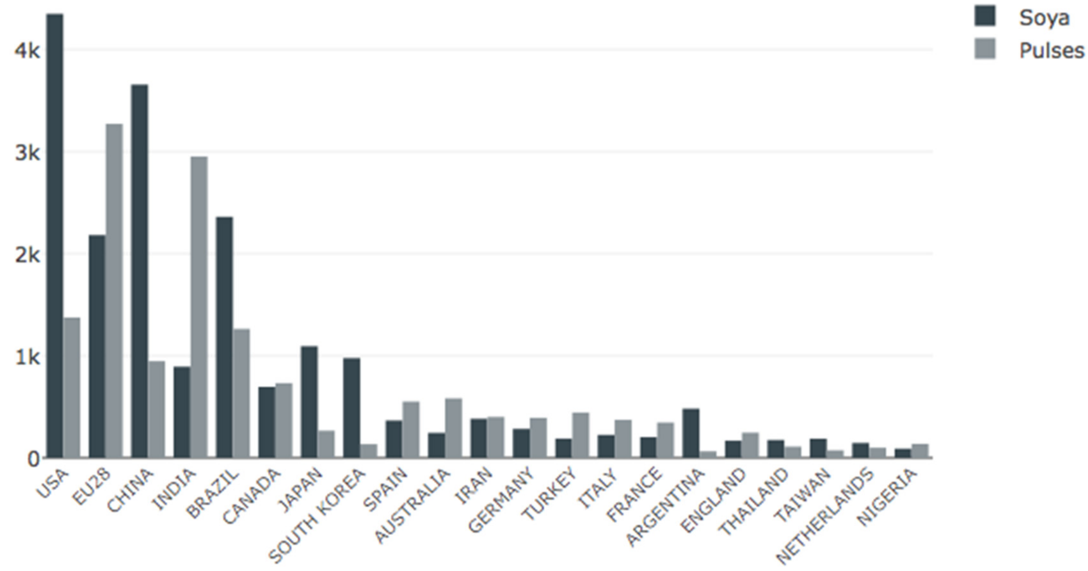
3.2.2. The Percentage of Soya and Pulses Records per Country Is Quite Stable over Time

Second, a clear opposition appears with some countries that focus more on pulses than on soya. This is particularly the case for the EU28 and India (and for Australia but with fewer records) and for the countries geographically close to them. Over the whole period, the EU28 accounted for more than a quarter of the studies on pulses, but less in the current decade because of India's increase. For instance, while the EU28 has two times fewer records in the current decade than during the previous period, it is the opposite for India which has doubled records in the current decade (Table 6). Canada is the only country publishing as much on soya as on pulses, whatever the period. Other countries work more on soya, increasing the imbalance with pulses: this is particularly the case for China which currently publishes nearly four times more on soya than on pulses, with the USA three times more, and Brazil twice more.

Among the 20 most publishing countries, there are six European countries in the current decade: Spain, Germany, France, Italy, England, and Poland. They also represent two-thirds of the records in the EU28. Figure 6a,b gives more details on the share of each European country in the EU28. One emerging trend is the increase of soya records, becoming more important than pulse records for the Netherlands and Romania, and near equal for Belgium, Denmark, and Austria. Poland, France, and the UK are countries with the highest proportion of pulse records compared to soya records. Overall, the ranking of European countries is stable over time, apart from the UK whose number of records is smaller in the current decade. Currently, Spain is the top European country on pulses regarding the number of publications, and it is also the top European country for pulse cultivation and consumption (Eurostats).

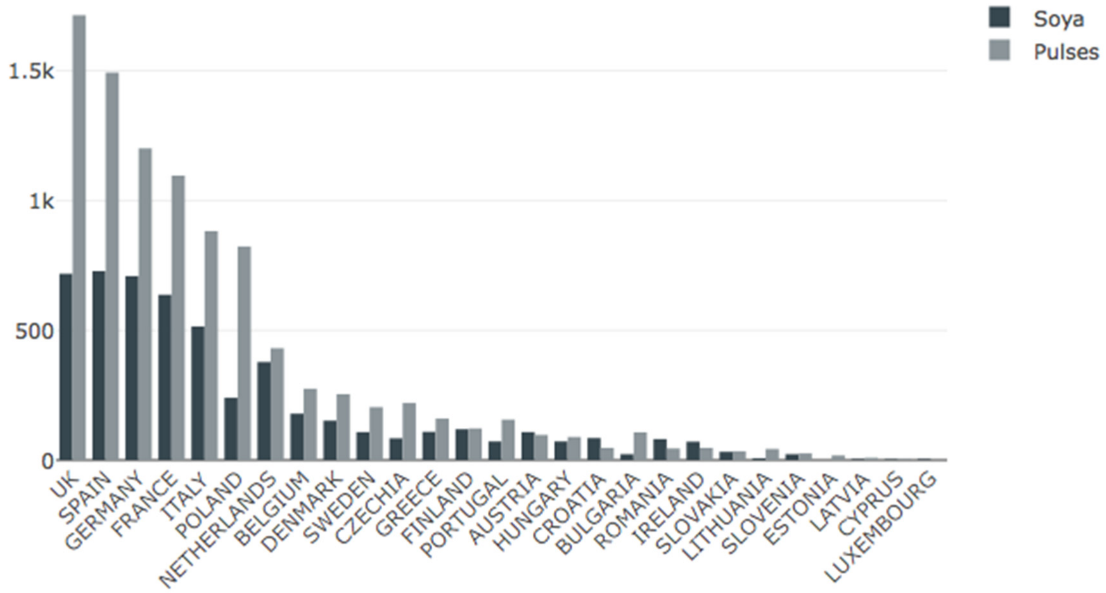


(**a**)

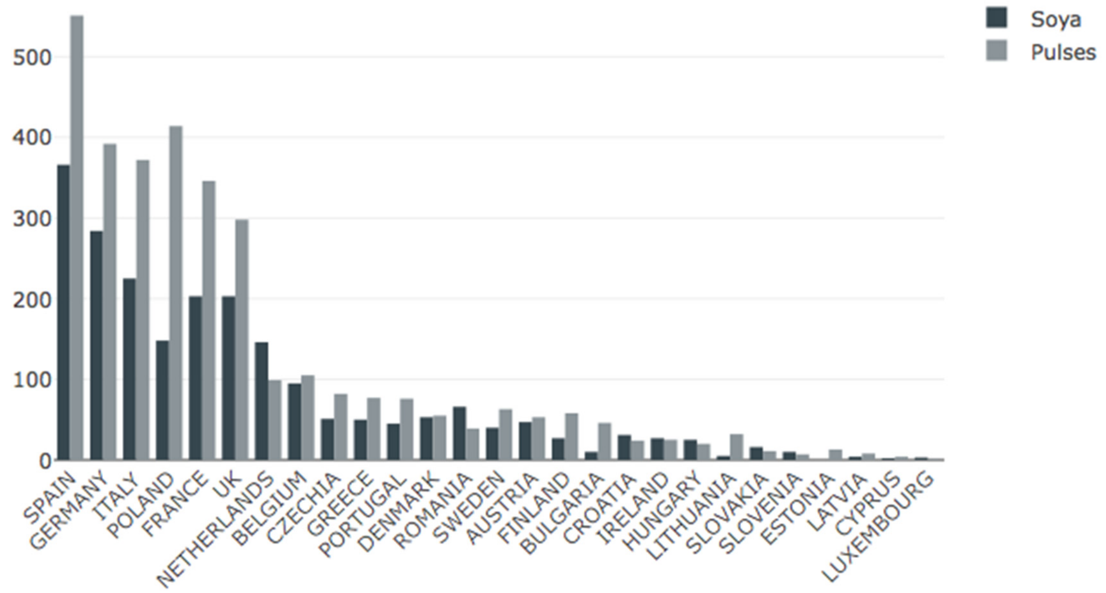


(**b**)

Figure 5. Soya and Pulses records across countries in the Fusion corpus. (**a**) 1980–2018. (**b**) 2010–2018. Note: the 20 highest frequencies are based on total records by country, a group count done for the EU28. Only records indexed with soya or with pulses were included (i.e., records co-indexed with several groups of grain-legumes were excluded). Proportional count for international collaboration records is applied. The country ranking is based on the total records number by country.



(**a**)



(**b**)

Figure 6. Soya and Pulses records across the EU28 countries in the Fusion corpus. (**a**) 1980–2018. (**b**) 2010–2018. Note: only records indexed with soya or with pulses were included (i.e., records co-indexed with several groups of grain-legumes were excluded). Proportional count for international collaboration records is applied. The country ranking is based on the total records number. Malta had no records.

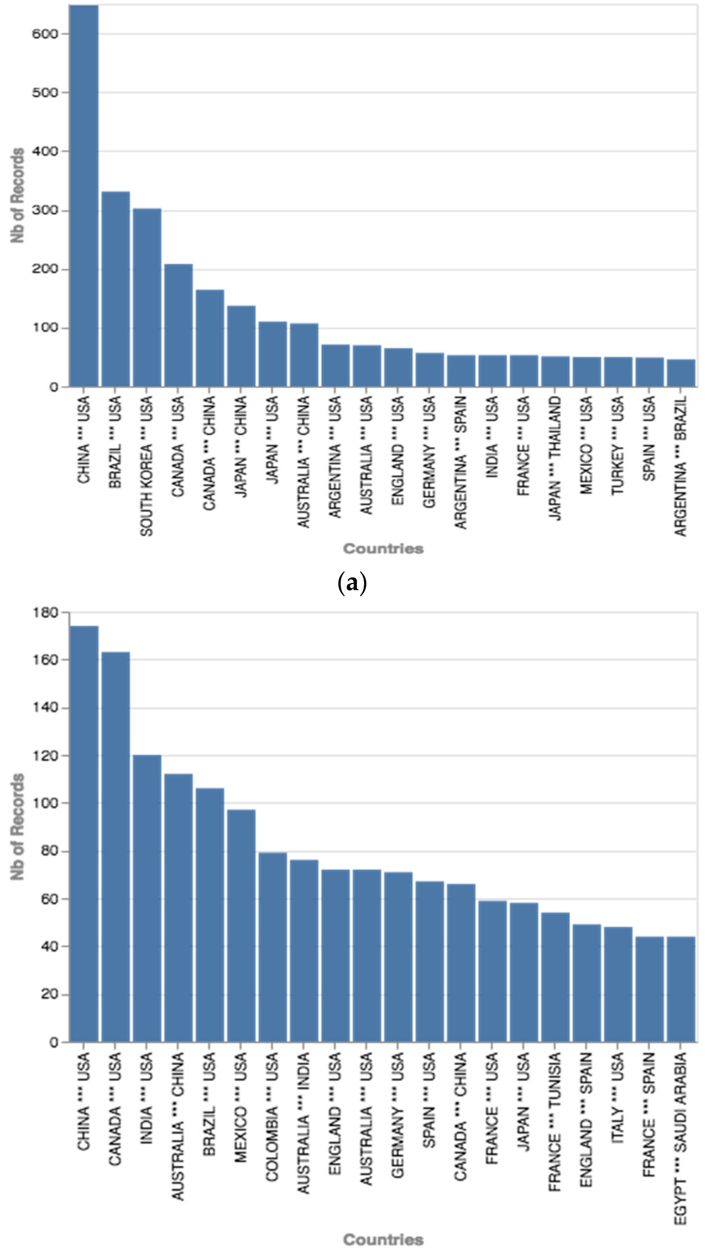
Table 6. The eight main areas/countries publishing the most on pulses or soya.

Area Rank *	Area or Country	1980–2018 Period						1980–2009 Period						2010–2018 Period					
		S / P	%	S	%	P	%	S	%	S	%	P	%	S / P	%	S	%	P	%
1	USA	18,238	24%	13,487	33%	4751	13%	12,524	30%	9148	42%	3376	17%	5714	16%	4339	22%	1375	9%
2	UE28	14,854	19%	5260	13%	9594	27%	9401	22%	3078	14%	6323	31%	5453	16%	2182	11%	3271	21%
4	INDIA	7043	9%	1634	4%	5409	15%	3198	8%	741	3%	2457	12%	3845	11%	893	5%	2952	19%
3	CHINA	5926	8%	4564	11%	1362	4%	1323	3%	908	4%	415	2%	4603	13%	3656	19%	947	6%
5	BRAZIL	5712	7%	3645	9%	2067	6%	2088	5%	1284	6%	804	4%	3624	10%	2361	12%	1263	8%
6	JAPAN	4286	6%	3147	8%	1139	3%	2926	7%	2053	9%	873	4%	1360	4%	1094	6%	266	2%
7	CANADA	3524	5%	1722	4%	1802	5%	2097	5%	1027	5%	1070	5%	1427	4%	695	4%	732	5%
8	AUSTRALIA	2530	3%	747	2%	1783	5%	1702	4%	502	2%	1200	6%	828	2%	245	1%	583	4%
SUBTOTAL **		62,113	80%	34,206	83%	27,907	78%	35,259	84%	18,741	86%	16,518	82%	26,854	76%	15,465	79%	11,389	73%
TOTAL ***		77,170	100%	41,413	100%	35,757	100%	42,014	100%	21,856	100%	20,158	100%	35,156	100%	19,557	100%	15,599	100%

Note: S means Soya; P means Pulses; S/P considers both. * Area ranking over 1980–2018 for Soya and Pulses records. ** SUBTOTAL corresponds to the sum of the eight main areas/countries. *** TOTAL corresponds to all countries. Number of records indexed either with soya (only) or with pulses (only), with proportional count for international collaboration records. These counts were applied on a subset of the FUSION corpus as authors' countries were not identified for nearly 13,000 records indexed with soya or pulses.

3.2.3. International Collaboration Research Is Increasing, but Unevenly on Soya or Pulses

The records involving several countries (international collaboration records) amount to 19% of the records indexed with soya or pulses (16% before 2010 and 23% in the current decade). We identified 3495 combinations of countries among those collaborations and most of those involved two countries. Figure 7a,b shows the most frequent collaborations on soya or pulses globally. We observed a clear leadership of the USA in those collaborations both for soya and pulses. Although there were fewer international records on pulses than on soya among the 20 most frequent collaborations, over the current decade 26% of records on pulses involved international collaboration compared with 22% for soya. This difference at a global scale is due to the increasing collaboration among the EU28 countries, which focus more on pulses than on soya. Over the current decade, this ranking of the most frequent collaborations has remained stable.



(**a**)

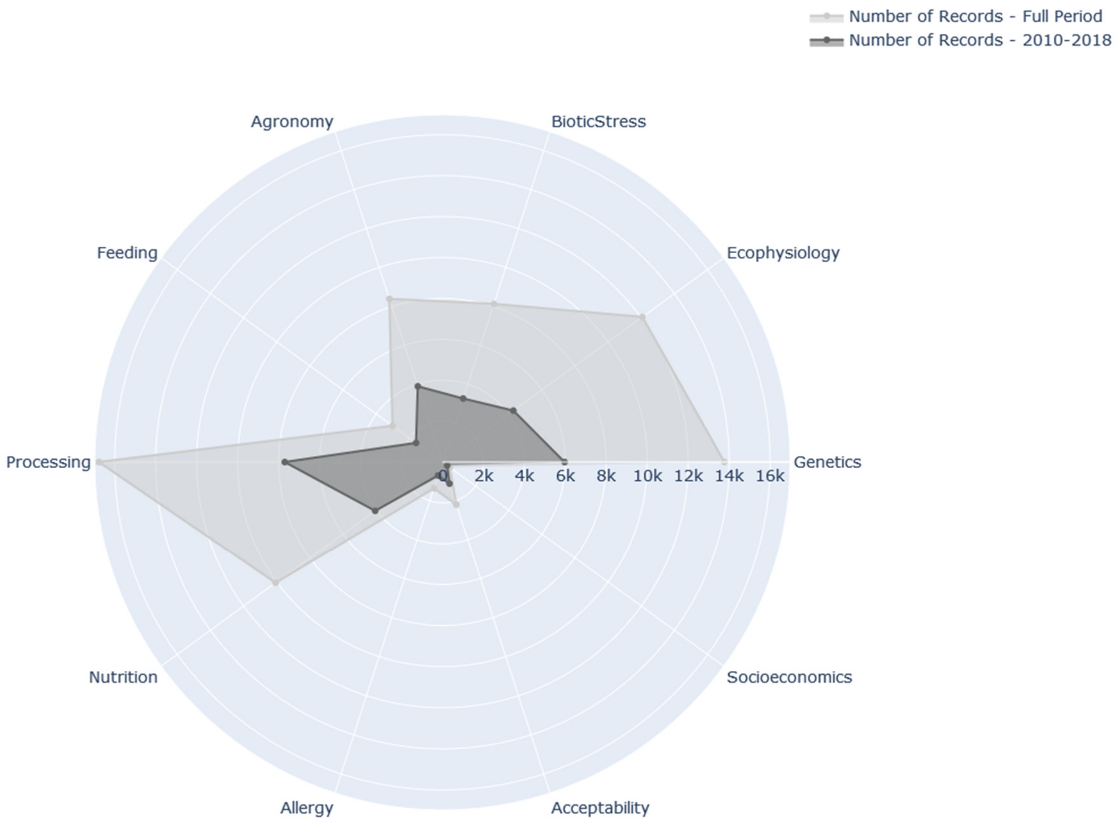
(**b**)

Figure 7. The 20 most frequent international collaborations 1980–2018 in the Fusion corpus. (**a**) Most frequent international collaborations on soya. (**b**) Most frequent international collaborations on pulses.

The counts in Figure 7a,b correspond to the number of records indexed with soya (only) and with pulses (only), respectively, and for which the authors were only from the countries specified in the horizontal axis. In the WoS, the country indexing is alphabetically ordered and does not correspond to the authors' order in records.

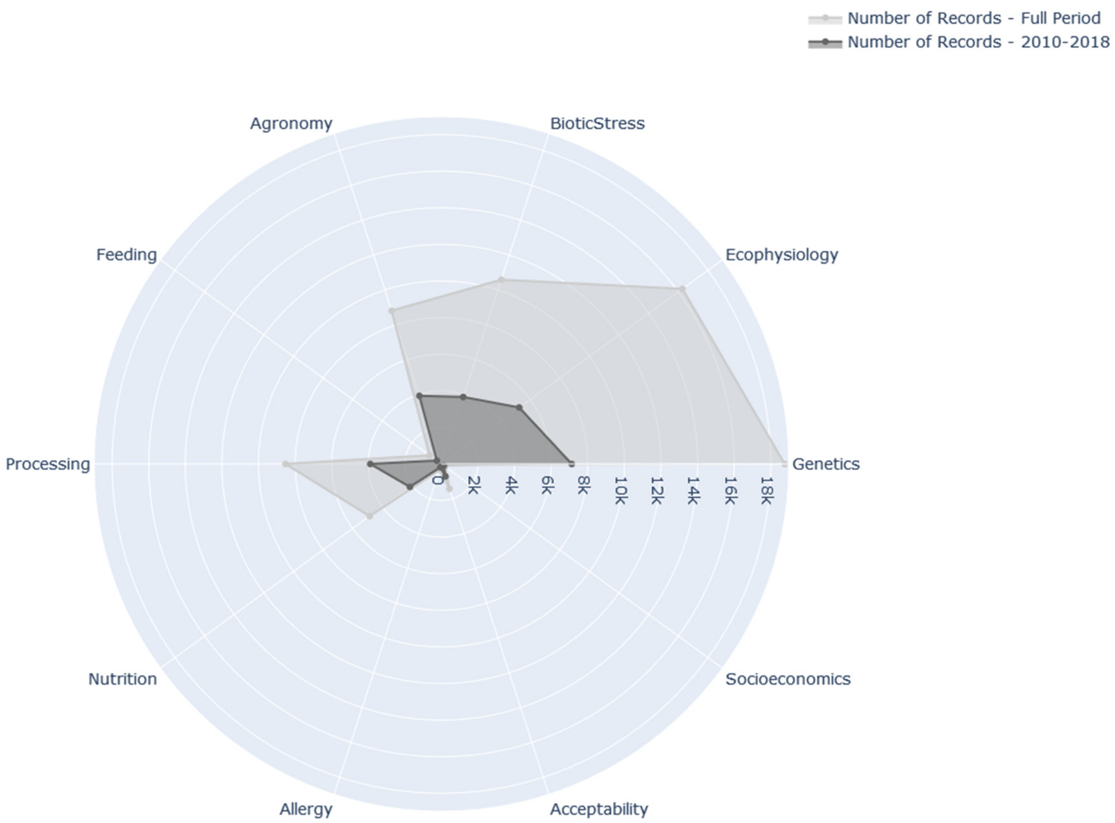
3.3. Percentage of Themes in the Soya and Pulses Literature

Figure 8 presents the shares of the 10 themes considered in the Fusion corpus for soya and pulses. First, the breakdown of themes differed for soya and pulses, even in recent years: Processing and Nutrition were more frequent themes for soya than for pulses. In addition, while total records for soya are slightly greater than for all pulses (see Section 3.1), "upstream" themes—such as Genetics, Ecophysiology, BioticStress and Agronomy—had a few more records for pulses than soya. These comparative figures clearly highlight the fact that "downstream" themes are less invested for pulses compared to soya. Moreover, as mentioned above, while feeding is a major outlet for soya, this theme did not represent a large share of the research on soya. However, records on feeding for soya double the number of records for pulses on the same theme. The remaining minor themes, such as Acceptability and Allergy, were more developed for soya than for pulses, even in the recent period.



(**a**)

Figure 8. *Cont.*



(**b**)

Figure 8. Shares of themes within records in the Fusion corpus over 1980–2018 and 2010–2018. (**a**) Records on soya. (**b**) Records on pulses.

The results show that the USA, the EU28, India, and China are the four main countries/areas publishing on soya or pulses (Table 6). We calculated the thematic shares for those main countries/areas compared to all other countries to illustrate the variation of themes investigated for pulses and soya (Figure A2). On average, while compared to the USA the EU28 globally focused more on pulses than on soya, for some themes, the gap between them is smaller. For instance, in Genetics and BioticStress, the USA worked on these themes for pulses as much as the EU28 did, while for soya the EU28 published little on these subjects.

Some themes have varying interest by countries. For instance, considering China's strong and recent focus on soya, the most investigated themes are, respectively: Processing, Genetics, Ecophysiology, Nutrition, and BioticStress. For the USA, these come in a different order: Genetics, Ecophysiology, BioticStress, Processing, and Agronomy.

We previously observed that "downstream" themes were less studied for pulses, yet this difference is less pronounced for the EU28 and India. For instance, they focused almost as much on soya as on pulses for Nutrition and Acceptability themes. Therefore, if spatial correlation between the scholarly records on a crop and its level of production in the country seems to be an explanation at first view, further investigation is needed to better understand the variation of themes between soya and pulses according to countries. This could be linked to the research strategies of a specific research team, which would need a socio-semantic network study to uncover fully.

As regards the Feeding theme, while increasing pulses in feed is an important goal for European public authorities that has led to millions in investments, there is still more research on soya by the EU28 than on pulses, representing about as many records as for the USA. Other countries are much involved in soya research for the Feeding theme, such as the considerable focus in South America (especially Brazil and Argentina).

3.4. Implications for Future Research Policy on Grain-Legumes

The shares of crops in research are strongly path-dependent, especially as regards soya and pulses. Soya dominates research on legumes at the global scale whatever the theme, being also the crop with a dominant market size compared to pulses (Appendix A, Table A1). In addition, we observed that the gap between soya and pulses has strongly grown over time for some countries such as China. For instance, when counting the records before 2010 for China, the gap between pulses and soya was not so great, but it strongly increased afterwards. In contrast, for other countries, the gap between soya and pulses has been stable: that is to say, no opposite trend appears between the two periods for any country or theme considered here. Globally, pulses benefit from less research activity than major crops such as soy, as advanced by other works [32]. As regards global plant-based protein markets for food, soya is mainly used in current product innovations compared to pulses [33]. Therefore, since markets do not seem to drive agrofood systems towards more agricultural diversity, this raises questions about how to shift research to provide knowledge that will make a transition towards more grain-legumes crop diversity possible. The present study did not examine research trends according to private or public funding. One possible avenue for future research would be to determine whether public funding favors diversity in agricultural research. Such an investigation would require analyzing the "Funding Acknowledgment Table" indexed by the WoS since 2008, through a list of institution names that need to be standardized for analysis. Finally, the shares of grain-legumes in research activity seems not to be totally dependent on the crop production of the countries and further investigation must be conducted to better analyze for which countries this spatial correlation is stronger.

As the share of research between soya and pulses is highly uneven, it is essential to increase research on pulses including a wide variety of species, compared to soya as a single species, as advocated during the IYP in 2016. One main challenge remains to create links between species, what some researchers call "translational" research. This would require important changes in research. For instance, if the large body of research on agronomy and ecophysiology were more systematically associated with crop modeling, their hypotheses and results would be easier to use for research protocols in other species. As shown by the relatively low number of studies combining at least two pulse species, there is also a need for more comparative analysis across species and for all the themes identified in this study. Research planning and policy should also consider that the model used for oil-legumes, focused on one species (soybean), will not work with pulses, which have strong specificities with regards to consumer preferences and food processing. Public investments and public–private partnerships will be essential to ensure that an increase in pulse research funding does not erode this pulse species diversity. It was outside the scope of this study to analyze financial investments in research on pulses, but the number of publications on these crops, their increase during the last two decades, and the identification of leading countries could support an international approach to the research on these crops. They would be essential components of nutrition sensitive agriculture in the Earth's margins, recently termed "environmental nutrition" [34].

Moreover, as future pulse development would be for food and nutrition, more investment in food sciences, including processing, is crucial for providing diverse pulse-based products that ensure food security and good health. While for upstream themes required to increase production (such as Genetics and Ecophysiology) pulses have received a good deal of attention, more research in food sciences is still needed to develop markets as argued in other studies. The results here show that both Nutrition and Processing themes were strong for soya, while the food outlet for whole-grain soya represents only around 5% of its production [9]. That is to say, around 15 million tons of soya production is entirely used for food, while the food outlet for pulses is considered to be (at least) around 50% of the global production, that is, accounting for a higher food outlet of around 50 million tons according to the volumes of those crops (Table A1). However, our study found less research for pulses in themes related to food sciences. By creating products meeting food habits and preferences, improving nutritional values, and increasing their usage by renowned food brands, research in processing could be a way to increase pulse consumption, to value under-used pulses species, and finally to build a value chain by making pulses more than a low-cost substitute for animal proteins [35]. A compromise between marketing opportunities and affordable healthy food for consumers should more often be an objective of research in food sciences. Agricultural and food sciences have to work hand in hand for sustainable agriculture, designing coupled innovations between the upstream and downstream of supply chains [36]. Therefore, more publications co-indexed with agricultural and food sciences are expected.

4. Conclusions

Understanding the dynamics of research is an epistemic project that concerns all sciences and is one primary focus of STS. Analyzing bibliometric datasets help both decision-makers for science policies and scholars to orient science and, in the end, innovation. By using scientometric indicators on the metadata of scholarly documents retrieved from the WoS, we gave an overview of the research output on grain-legumes since 1980 at the global scale. We quantified the shares of species, analyzed by main academic fields (i.e., themes) and countries. To establish these results, this work developed a rigorous methodology for building a bibliometric dataset and a processing software, which can be used for further bibliometric research on other crops to bring a larger analysis of the dynamics of research activity on agricultural and food systems. We first discuss main implications of this work, and then propose further works to be conducted to improve our analyses.

First, these results show that research on grain-legumes is path-dependent and is strongly linked to the size of their agricultural production, as the main species researched in the past continue to be the main focus of research today. These findings should foster further discussion and reflection among the scientific community about grain-legumes, especially regarding the challenge of greater crop diversity. Above all, researchers must be encouraged to create links among species to enlarge the application of their results, as we found few records mentioning several species together. In addition, while upstream themes have received considerable attention (such as GENETICS and ECOPHYSIOLOGY), downstream themes have received unequal interest among crops; particularly, compared to soya, pulses have been less researched in relation to downstream themes of food sciences. This imbalance questions the capacity of research to develop knowledge that would enable more food outlets for pulses, as expected by the United Nations during the IYP. Finally, the results show that most research on grain-legumes has been conducted by eight countries/geographical areas (including the European Union). While some geographical areas have done more research on pulses than on soya, such as the EU28 and India, others specialized on soya, such as many American countries. Newcomers such as China clearly made the choice to invest more in research on soya, establishing important collaborations with the USA.

Second, there is a strategic interest in bibliometric studies on the way research is conducted on agricultural and agrofood systems, in order to define new research priorities. This type of work is not an easy task, and collaboration between scientific experts on the themes investigated and experts in scientometrics is essential. Given that we are faced with an ever increasing amount of published data, involving researchers on the measurement of sciences is essential. Moreover, those researchers could advise other researchers on the way to communicate about their work. Indeed, the way bibliometric datasets are collected and how researchers describe their works through the title, abstract, and keywords of records impact the ways we can analyze research patterns through bibliometric data. Title, abstract, and keywords are the main data on which scientometric studies relies, and thus their contents strongly determine the results obtained. Our findings raise another question about the way species are mentioned: a common dictionary of species names in scientific platforms, such as the WoS, would help to better follow species, and to conduct longitudinal analysis on the percentage of any crop in research and their links.

Thirdly, we used English-language scholarly publications, such as peer-reviewed articles, books, and book chapters, to reflect research activity and to give an overview of the majority of the scientific knowledge on grain-legumes. Other data collections could enrich our dataset to have a more complete overview. The first improvement will be to add records retrieved from the Scopus Collection (the other most used bibliometric platform), even though the WoS and Scopus overlap considerably for certain disciplines [26]. This could lead to some adjustments in the shares of themes, but probably not in their ranking. Other enrichments such as using alternative search engines, e.g., Google Scholar (GS), will require methodological solutions. GS provides access to larger collection of research such as PhD theses, scientific reports, and non-peer-reviewed articles, but no software exists to retrieve large collections from GS. In addition, no rigorous metadata are associated to enable relevant analysis among the various documents once they are collected. Last, while English dominates scientific publications, other languages are still used and including them would complicate larger record harvesting and analyses from GS. "The second most frequent language of unique GS citations was Chinese (4–12%), and all other languages have a share of 4% or lower across all subject areas. A few (5–10%) unique GS citations were published in languages outside the top 11 most frequently used languages overall" [26] (p. 14).

To have a better overview of the evolution of this scholarly scientific knowledge on grain-legumes, further studies can be done (based on this dataset or an enriched dataset) such as semantic or socio-semantic network analysis. Co-words and institution mapping would provide overviews of the main and minor concepts and knowledge that characterize research according to species, themes, and countries. By identifying the main relationships between terms, bridges of knowledge and new research areas can be identified within and among species. In particular, as grain-legumes are recognized to have a high potential for more sustainable agriculture, it would be interesting to analyze the development of specific targets and wordings aiming at increasing the role of these species in a sustainable agricultural production, for both food and feed.

Lastly, understanding the relationships between markets trends and research directions remains a main challenge of the STS. Future works could analyze the spatial correlation between scholarly documents on crops and the crops production/consumption levels within countries to analyze links between research activity and societal demands, as recently conducted on rice species [37]. Considering also the metadata associated to funding in notices (indexed in the WoS since 2008) will allow evaluating the shares of independent research (i.e., non-commercial funding) compared to other private-based funding, and could reveal different research trajectories regarding the specialization vs. diversification of crops under research.

Author Contributions: Conceptualization, M.-B.M. and G.C.; Data curation, M.-B.M., G.C., M.L., H.L. and D.M.; Formal analysis, M.-B.M., G.C., M.L., M.-J.A., M.A., G.A., A.B., L.B., J.-M.C., G.D., M.-H.J., E.-P.J., H.J., C.L., H.L., V.M., D.M., M.-L.P.-N., C.N.-T., T.S., A.-S.V., S.W. and J.W.; Funding acquisition, M.-B.M.; Investigation, M.-B.M., G.C. and M.L.; Methodology, M.-B.M., G.C., M.L., M.-J.A., M.A., G.A., A.B., L.B., J.-M.C., G.D., M.-H.J., E.-P.J., H.J., C.L., H.L., V.M., D.M., M.-L.P.-N., C.N.-T., T.S., A.-S.V., S.W. and J.W.; Project administration, M.-B.M.; Resources, M.-B.M. and G.C.; Software, G.C.; Supervision, M.-B.M.; Validation, M.-B.M., G.C., M.L., G.P., M.-J.A., M.A., G.A., A.B., L.B., J.-M.C., G.D., M.-H.J., E.-P.J., H.J., C.L., H.L., V.M., D.M., M.-L.P.-N., C.N.-T., A.-S.V., S.W. and J.W.; Visualization, M.-B.M., G.C., H.L. and T.S.; Writing—original draft, M.-B.M., G.C. and G.P.; and Writing—review and editing, M.-B.M., G.C., M.L., G.P., M.-J.A., M.A., G.A., A.B., L.B., J.-M.C., G.D., M.-H.J., E.-P.J., H.J., C.L., H.L., V.M., D.M., M.-L.P.-N., C.N.-T., A.-S.V., S.W. and J.W.

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 by the French National Institute for Agricultural Research (INRA).

Acknowledgments: The authors acknowledge the two anonymous reviewers whose comments and suggestions helped sharpen the argument, and Cynthia J. Johnson for her helpful comments and English editing of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Global production of main pulses, soya and cereals, trienniums ending 1971, 1981, 1991, 2001, 2011, and 2017 (million metric tons). Source: FAOstat.

Species	Year					
	1971	**1981**	**1991**	**2001**	**2011**	**2017**
Bean (dry)	12	15	18	18	24	34
Chickpea	6	5	8	6	11	14
Pea (dry)	9	7	12	10	10	16
Faba/broadbean	8	8	6	8	8	8
Lentil	1	1	2	3	4	7
Pigeon pea	2	2	2	3	4	6
Cowpea	1	1	2	3	4	7
Vetches	1	1	1	1	1	1
Lupin	0.3	1	1	1	1	1
Bambara Bean	0.03	0.03	0.08	0.08	0.14	0.18
Other pulses	3	2	4	3	3	4
Total Pulses	42	41	56	56	69	95
Soya	45	88	102	177	261	352
Cereals	1229	1632	1890	2104	2588	2980

Table A2. Links to the 11 WoS queries designed to delineate the corpus under study, each query capturing thematic corpus on grain-legumes (e.g., Allergy and BioticStress). The Species query was combined to thematic queries or used isolated.

Query Name	Link to the Search Query Applied on the Web of Science (WoS) Core Collection
THE SPECIES SEARCH QUERY	
SPECIES	DUC, Gérard; WERY, Jacques; MAGRINI, Marie-Benoit; CABANAC, Guillaume, 2019, “Grain-Legumes Species WoS DataSet 1980–2018”, https://doi.org/10.15454/QBQFCX, Portail Data Inra
THE 10 THEMATIC SEARCH QUERIES	
GENETICS	DUC, Gérard; WERY, Jacques; MILLOT, Dominique; CABANAC, Guillaume, 2019, “Genetics and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/PFV9JK, Portail Data Inra
AGRONOMY	JEUFFROY, Marie-Hélène; BEDOUSSAC, Laurent; MILLOT, Dominique; CABANAC, Guillaume, 2019, “Agronomy and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/W6BAUG, Portail Data Inra
ECOPHYSIOLOGY	VOISIN, Anne-Sophie; JOURNET, Etienne-Pascal; LEISER, Hugues; CABANAC, Guillaume, 2019, “Ecophysiology and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/F0CNNS, Portail Data Inra
BIOTICSTRESS	BARANGER, Alain; PILET-NAYEL, Marie-Laure; MILLOT, Dominique; CABANAC, Guillaume, 2019, “Biotic Stress and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/L79X2K, Portail Data Inra
FEEDING	JUIN, Hervé; LEISER, Hugues; CABANAC, Guillaume, 2019, “Feeding and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/BNKFVC, Portail Data Inra
PROCESSING	ANTON, Marc; MICARD, Valérie; NGUYEN-THE, Christophe; LEISER, Hugues; CABANAC, Guillaume, 2019, “Processing and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/VP7PRI, Portail Data Inra
NUTRITION	AMIOT-CARLIN, Marie-Josephe; CHARDIGNY, Jean-Michel; WALRAND, Stéphane; LEISER, Hugues; CABANAC, Guillaume, 2019, “Nutrition and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/5MI04S, Portail Data Inra
ALLERGY	LARRE, Colette; DENERY, Sandrine; LESIER, Hugues; CABANAC, Guillaume, 2019, “Allergy and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/BZG0R7, Portail Data Inra
ACCEPTABILITY	ARVISENET, Gaelle; MAGRINI, Marie-Benoit; LEISER, Hugues; CABANAC, Guillaume, 2019, “Acceptability and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/PDXRYM, Portail Data Inra
SOCIOECONOMICS	Magrini, Marie-Benoit; Plumecocq, Gael; Leiser, Hugues; Cabanac, Guillaume, 2019, “Socioeconomics and Grain-Legumes WoS DataSet 1980–2018”, https://doi.org/10.15454/JNIPX5, Portail Data Inra

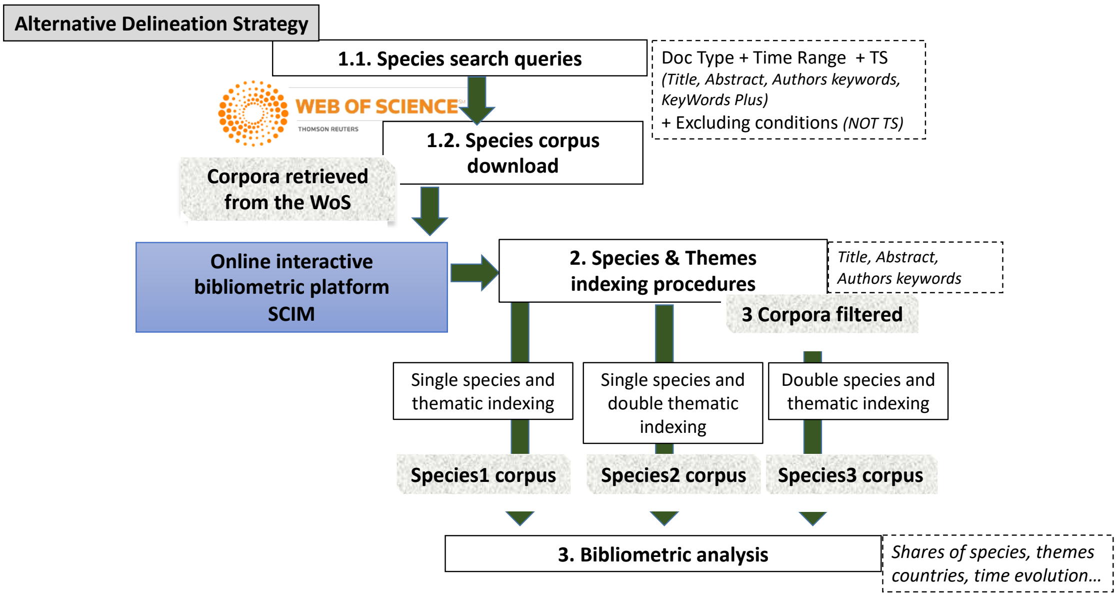
Table A3. Breakdown of the Fusion corpus in the 10 underlying themes (single or combined themes indexed).

Themes Index Colum A	**Records Number Indexed with a Single Theme ***	**Frequency Ranking of the Single Theme**	**Other Themes Index Combined with the Theme in Colum A ****													
			1st Most Freq.	**Nb. Records**	**2nd Most Freq.**	**Nb. Records**	**3nd Most Freq.**	**Nb. Records**	**4th Most Freq.**	**Nb. Records**	**5th Most Freq.**	**Nb. Records**	**6th Most Freq.**	**Nb. Records**	**Total nb of Records Indexed with the Theme of Column A**	**Share of Themes**
Genetics	13,336	1	Ecophy.	6845	BioticS.	4887	Agro.	2176	Agro. & Ecophy.	1250	Ecophy. & BioticS.	1119	Process.	977	34,388	22%
Ecophysiology (Ecophy.)	12,841	2	Genetics	6845	Agro.	3263	BioticS.s	1411	Agro. & Genetics	1250	BioticS. & Genetics	1119	Process.	878	29,867	19%
Processing (Process.)	12,049	3	Nutrition	5965	Accep.	1298	Nutrition & Accept.	1046	Allergy	1015	Genetics	977	Ecophy.	878	27,180	17%
BioticStress (BioticS.)	9109	4	Genetics	4887	Ecophy.	1411	Agro.	1556	Ecophy. & Genetics	1119	Accept. & Nutrition	1046	Agro. & Genetics	583	20,243	13%
Agronomy (Agro.)	7982	5	Ecophy.	3263	Genetics	2176	BioticS.	1556	Ecophy. & Genetics	1250	BioticS. & Genetics	583	Process.	505	19,070	12%
Nutrition	4866	9	Process.	5965	Accept. & Process.	1046	Genetics & Process.	606	Genetics	578	Ecophy.	238	Ecophy. & Process.	214	15,383	10%
Feeding	2888	11	Genetics	181	Process.	147	Nutrition	113	Nutrition & Process.	80	Agro.	42	Ecophysiology	36	3654	2%
Allergy	1891	13	Process.	1015	Nutrition & Process.	191	Nutrition	148	Genetics	59	Genetics & Process.	28	Genetics & Nutrition	27	3650	2%
Acceptability (Accept.)	745	23	Process.	1298	Nutrition & Process.	1046	Nutrition	194	Genetics & Process.	105	Genetics	90	Genetics & Nutrition	81	3972	3%
Socioeconomics (Socioeco.)	565	27	Agro.	125	Process.	44	Genetics	33	Agro. & Ecophy.	31	Agro. & Genetics	27	Accept. & Process.	15	963	1%
SUBTOTAL	66,272														158,370	100%
in % of corpus	61%															
Total Corpus	107,823															

Note: * that is neither of the Title, Abstract or Authors' keywords contain terms present in the query of any other theme. ** that is the Title, Abstract or authors' keywords contain terms linked to those two themes, but not of other themes. All the various combination of frequencies presented in this table represent 99% of the records in Fusion corpus.

Appendix B. Robustness Assessment of the Delineation Process: Testing an Alternative Strategy

We experimented with an alternative search strategy using the WoS to delineate a core scientific literature dataset on grain-legumes. As explained in Section 2, to obtain close links between grain-legumes species and specific thematic terms, we gave preference to the use of the Boolean character NEAR/10 in the search query design of most themes. In that way, the search query matched records whose terms joined by the operator were within 10 words of each other. However, to appreciate the differences in the frequency of records retrieved, depending on the use of the operator NEAR or not, we also built other bibliometric datasets on those grain-legumes, resulting from search queries without the NEAR operator. This other strategy led to the so-called SPECIES1, SPECIES2, and SPECIES3 corpora, following an alternative methodology illustrated in Figure A1. This alternative delineation strategy used the same terms in the search queries than the one exposed in Section 2.



Alternative Delineation Strategy
1.1. Species search queries
Doc Type + Time Range + TS
(Title, Abstract, Authors keywords, KeyWords Plus)
+ Excluding conditions (NOT TS)
WEB OF SCIENCE
THOMSON REUTERS
1.2. Species corpus download
Corpora retrieved from the WoS
Online interactive bibliometric platform SCIM
2. Species & Themes indexing procedures
Title, Abstract, Authors keywords
3 Corpora filtered
Single species and thematic indexing
Single species and double thematic indexing
Double species and thematic indexing
Species1 corpus
Species2 corpus
Species3 corpus
3. Bibliometric analysis
Shares of species, themes countries, time evolution...

Figure A1. Main steps to build the bibliometric dataset according to an alternative delineation strategy. Figure A1 summarizes the main steps followed to build the corpora quite similar to that of Figure 1 (Section 2), but with a change in the way to design search queries and the indexing procedure. This alternative delineation strategy led to three other corpora called SPECIES1, SPECIES2, and SPECIES3.

SPECIES1, SPECIES2, and SPECIES3 are the corpora retrieved from the WoS by using the SPECIES search query only, on which various indexing step strategies were applied. The aforementioned indexing step consists in keeping the records having a term of the search query among the *Title, Abstract or Author's keywords* only (i.e., filtered from the *KeyWord Plus*). This also led to indexing each record with the search terms found in *Title, Abstract* or *Author's keywords*. First, each record was indexed with one or several legume species according to the SPECIES terms occurring in the record (see Table 3 for species indexation). Second, we indexed each record relatively to the 10 thematic subjects when any of the terms of the records matched the terms of the thematic query (Appendix A, Table A2). In other words, this strategy did not rely on the NEAR operator but on the indexing procedure applied on the SPECIES corpus downloaded from the WoS. The three variants of the SPECIES corpora were built, depending on the way the indexing procedure was applied:

- SPECIES1 has single species and thematic indexing. A record was indexed with a species term and with a thematic corpus, if at least one term of the species query and at least one term of the thematic query occurred in the record.

- Species2 has single species indexing and double thematic indexing. A record was indexed with a species term and with a thematic corpus, if at least one term of the species query and at least two terms of the thematic query occurred in the record.
- Species3 has both double species and thematic indexing. A record was indexed with a species term and with a thematic corpus, if at least two terms of the species query and at least two terms of the thematic query occurred in the record.

The size of these corpora varied and are reported in Table A4.

We observed that, of course, the more we restricted the strategy on search query and indexing, the fewer records were included. We matched the records between Fusion and Species3 corpora: we observed a difference of 22,661 records caught by Species3 but not present in Fusion; inversely, 27,813 records were caught by Fusion and not present in Species3.

However, above all, the correspondence between the wcs and the themes defined by experts was less adequate for the Species1, Species2, and Species3 corpora compared to the Fusion corpus. For instance, the classification of the records with the 10 themes investigated by experts was not really relevant in Species3, while we observed strong correspondence between the wcs and the thematic indexes in Fusion. Therefore, the forthcoming analysis by themes is more biased in Species3, as this corpus induces a biased representation of the 10 themes compared with Fusion. This was encountered, for instance, with the Nutrition theme that included a lot of records dealing with plant growth and not with human nutrition in Species3.

For evidence on the stronger relevance of the delineation strategy kept for Fusion Corpus, compared to the alternative delineation strategy, we present Table A5 should the number of records according to thematic indexing, respectively, in the Fusion and Species3 corpora. We observed that in Fusion, the records indexed with a single theme were more frequent (61%) than in Species3 (25%). Hence, the Fusion corpus led to a quite clear thematic classification of the records among the 10 themes investigated, given the fact that among the remaining co-indexed thematic records 30% were indexed with two themes, and 8% with three themes. More precisely when considering thematic ranking, in Fusion corpus, the five first thematic indexes frequencies (concerning records indexed with only one theme) were: Genetics, Ecophysiology, Processing, BioticStress and Agronomy. The remaining single theme indexes (Nutrition, Feeding, Allergy, Acceptability, and Socioeconomics) appear with less frequency as there are many fewer records on these themes, and that Genetics, Ecophysiology, Processing, BioticStress, and Agronomy are themes whose co-indexing between themselves has a high frequency of records (on that point, see Table A3 that presents frequencies on co-indexing themes in the Fusion corpus). In all, these five themes were the most frequent, respectively, 22%, 19%, 17%, 13%, and 12%.

All these remarks show the importance of the delineation strategy in building a bibliometric corpus. Moreover, it is clear that conducting bibliometric analysis requires considerable support of experts in the scientific themes investigated, since relying on wcs alone is not sufficient to delineate a relevant bibliometric corpus. Consequently, for us, the methodology kept to establish Fusion corpus is the most appropriate for identifying a "core" literature dataset on grain-legumes whose records can be classified by relevant themes (see also Table A3). Notwithstanding, some statistics presented in the Section 3, were also calculated on the datasets Species1, -2, and -3, to be compared to the ones established on Fusion corpus. In particular we observed that the shares of species were similar regardless of the delineation strategy of the bibliometric corpora (Table A6).

Table A4. Size of the bibliometric corpora depending on the delineation strategy.

	Delineation Strategy Kept			Alternative Delineation Strategy		
	Corpus Fusion			Corpus Species1	Corpus Species2	Corpus Species3
Search query applied on the WoS	For most thematic corpora, species and thematic terms combined with operator NEAR/10. See the ten thematic search queries in Appendix A.			Species terms only. Species search query in Appendix A.		
Excluding conditions	Some terms restrictions and WCS restrictions, depending on the thematic corpus.			The same term restrictions as in Fusion, but no WCS restrictions.		
Number of records retrieved from the WoS per corpus	Genetics	34,968		202,144		
	Agronomy	19,427				
	Ecophysiology	30,365				
	BioticStress	20,853				
	Feeding	4336				
	Processing	35,754				
	Nutrition	16,863				
	Allergy	5435				
	Acceptability	5459				
	Socioeconomics	1431				
Indexing procedure	One occurrence in the species terms and one occurrence in the thematic terms.			One occurrence in the species terms and one occurrence in the thematic terms.	One occurrence in the species terms and two occurrences in the thematic terms.	Two occurrences in the species terms and two occurrences in the thematic terms.
Number of records kept after indexing (share of records kept in %)	Genetics	34,388	98%	160,238 (79%)	142,763 (71%)	100,248 (50%)
	Agronomy	19,070	98%			
	Ecophysiology	29,867	98%			
	BioticStress	20,243	97%			
	Feeding	3654	84%			
	Processing	27,180	76%			
	Nutrition	15,383	91%			
	Allergy	3650	67%			
	Acceptability	3972	73%			
	Socioeconomics	963	67%			
Final number of records	Thematic corpora merged without duplicates: 107,823			160,238	142,763	100,248

Table A5. Breakdown of the Fusion and Species3 corpora into the 10 underlying themes.

FUSION Corpus					SPECIES3 Corpus				
Themes Index Colum A	**Records Number Indexed with a Single Theme ***	**Frequency Ranking of the Single Theme**	**Records Number Indexed with the Theme of Column A**	**Share of Themes**	**Themes Index Colum A**	**Records Number Indexed with a Single Theme ***	**Frequency Ranking of the Single Theme**	**Records Number Indexed with the Theme of Column A**	**Share of Themes**
Genetics	13,336	34,388	34,388	22%	Genetics	35,281	35,281	35,281	15%
Ecophy.	12,841	29,867	29,867	19%	Ecophy.	48,889	48,889	48,889	20%
Process.	12,049	27,180	27,180	17%	Process.	30,511	30,511	30,511	13%
BioticSt.	9109	20,243	20,243	13%	BioticSt.	17,839	17,839	17,839	7%
Agronomy	7982	19,070	19,070	12%	Agronomy	16,789	16,789	16,789	7%
Nutrition	4866	15,383	15,383	10%	Nutrition	56,021	56,021	56,021	23%
Feeding	2888	3654	3654	2%	Feeding	26,296	26,296	26,296	11%
Allergy	1891	3650	3650	2%	Allergy	2157	2157	2157	1%
Accept.	745	3972	3972	3%	Accept.	2280	2280	2280	1%
Socioeco.	565	963	963	1%	Socioeco.	3427	3427	3427	1%
SUBTOTAL	66,272		158 370	100%	158 370	24,946		239 490	100%
in %	61%				in %	25%			
Total FUSION Corpus	107 823				Total SPECIES3 Corpus	100 248			

Note: * A term keyword not contained in the title, abstract or authors' terms present in the query of any other theme.

Table A6. Number and share of the records in Species1, -2, and -3 corpora related to soya and pulses, groundnut and lathyrus-vicia broken down by periods.

Family species index terms of records	**SPECIES3**			**SPECIES2**			**SPECIES1**		
	1980–2018	**1980–1999**	**2000–2018**	**1980–2018**	**1980–1999**	**2000–2018**	**1980–2018**	**1980–1999**	**2000–2018**
Only Soya—G1	42,861	9372	33,489	62,233	15,330	46,903	68,790	19,750	49,040
Soya—G1—and a generic term *	533	108	425	1840	346	1494	1895	360	1535
Only other Pulses than PFL—G3	21,394	5401	15,993	25,268	7765	17,503	28,785	10,431	18,354
Other Pulses than PFL—G3—and a generic term	16,755	5625	11,130	22,447	8473	13,974	26,619	11,412	15,207
Only PFL—G2	1736	308	1428	3851	684	3167	3966	745	3221
PFL—G2—and a generic term	1303	253	1050	3111	656	2455	3227	689	2538
Subtotal Soya/Pulses	84,582	21,067	63,515	118,750	33,254	85,496	133,282	43,387	89,895
% Soya in Soya/Pulses subtotal	51.3%	45.0%	53.4%	54.0%	47.1%	56.6%	53.0%	46.4%	56.3%
% Soya/Pulses in corpus period	84.4%	85.4%	84.0%	83.2%	84.8%	82.6%	83.2%	85.2%	82.2%
Groundnut	8491	1900	6591	12,133	2883	9250	14,268	4055	10,213
% Groundnut in corpus period	8.5%	7.7%	8.7%	8.5%	7.4%	8.9%	8.9%	8.0%	9.3%
Lathyrus or Vicia	1374	310	1064	1684	427	1257	1907	539	1368
% Lathyrus/Vicia in corpus period	1.4%	1.3%	1.4%	1.2%	1.1%	1.2%	1.2%	1.1%	1.3%
Corpus total for the period	100,248	24,661	75,587	142,763	39,216	103,547	160,238	50,929	109,309
% period in corpus total 1980–2018	100%	25%	75%	100%	27%	73%	100%	32%	68%
Generic term only	1745	364	1381	1720	393	1327	1804	412	1392

Note: Each line in the table are excluding count from each other. PFL: pea, fababean or lupin. * Lecture: For instance, this line reports the number of records containing only a generic term and a term linked to soya in title, abstract or authors' keywords for a generic term such as legumes.

Appendix C. A Brief Overview of Grain-Legumes Domestication and Their Development

Annual legumes (Papilionaceae/Fabaceae or Leguminosae) cultivated for their seeds are frequent companions of cereals in most parts of the world. They are attractive because, contrary to other flower plants, legumes can fix atmospheric nitrogen thanks to a symbiosis with root bacteria called Rhizobium. The cultivation of legumes helps enrich the soil with nitrogen; hence, cultivated with cereals in rotation or in association, they contribute to higher fertility in soils. Moreover, as they are rich in protein, they complement cereals in diets. This protein complementarity, allowing them to substitute for animal-based proteins, was essential for the development of traditional farming communities. As underlined in [30], each important civilization in history had their basic cereals with their companion legumes. For instance, in Western Asia and Europe, wheats and barleys were frequently cultivated with peas, lentils, chickpeas, and faba beans, while, maize with several species of beans (Phaseolus) in the Americas, more with groundnut in South America. In Africa, mil and sorgho were grown along with niebe and voandzou. Soya was added among cereals in China, lablab and mungo in India, etc.

Appendix C.1. First Signs before the Common Era

The first definite signs of domesticated plants in the Old World appeared around 10,500–10,100 years before the common era (BCE). Legumes appear frequently with cereals (wheat and barley): "several grain legumes appear as constant companions of the cereals" [30] (p.1). The most frequent pulses in the early Neolithic period in the Middle East (near the Mediterranean Basin) are lentils (*Lens culinaris*) and peas (*Pisum sativum*), and two more local legume crops are chickpeas (*Cicer arietinum*) and bitter vetches (*Vicia ervilia*). Additional legumes were cultivated later, such as grass peas (*Lathyrus sativus*) with some evidence in Greece and Bulgaria around 8000–7000 BCE. The origin and early spread of faba beans (*Vicia faba*) is less clear.

These archaeological findings reveal "a rule, not a single crop but rather a combination of cereals, pulses, and flax appears in these early farming villages. Moreover, the assemblage seems to be similar throughout the Fertile Crescent. In other words, a common package of grain crops characterizes the development of agriculture in this 'core area'" [30] (p. 2–3) on the spread of those crops over Europe and Mediterranean Basin during 10,500–20,000 BCE.

In the Mediterranean basin and Europe, evidence from the beginning of the Common Era has found on the cultivation of the fenugreeks and the lupins, as well as the grass pea and vicias. As regards the Mesoamerican area, an illustration of legumes development was the "three sisters" system (maize, beans, and squash intercropping).

Appendix C.2. Antiquity Period

Legumes with bigger grains such as we know today reached this size in the Antiquity period. Domestication of these plants brought several major changes in plant architecture, pod size, and lodging resistance. Through cultivation, the stems became sturdier and stiffer, and had a reduced propensity to climb, to make them more easily cultivable in the field. Some wild type chemical defenses had also been counter-selected to favor their consumption. Many wild legumes contain strong toxins and antimetabolites that protect them from animal predation. Gradually, the techniques of cooking and soaking or fermentation allowed the seeds to be healthy for consumption.

Roman texts report different frequencies of legumes use for human consumption. Lentils, peas, and faba beans were the most widely consumed throughout Europe and the Middle East, as well as chickpeas further south. Fenugreek was used as a condiment, especially in the Mediterranean Basin. Lupines were more concentrated around Greece and Egypt, and seems to have been domesticated later in Antiquity. Some species within the genus *Vicia*, having a particular taste, were consumed only during periods of famine and were reserved for medicinal purposes. Certain preferences were established for each region, still marking our culinary traditions today. For example, beans were more heavily consumed in Egypt and Spain, lentils in France.

Appendix C.3. From the Middle Ages to the Modern Period

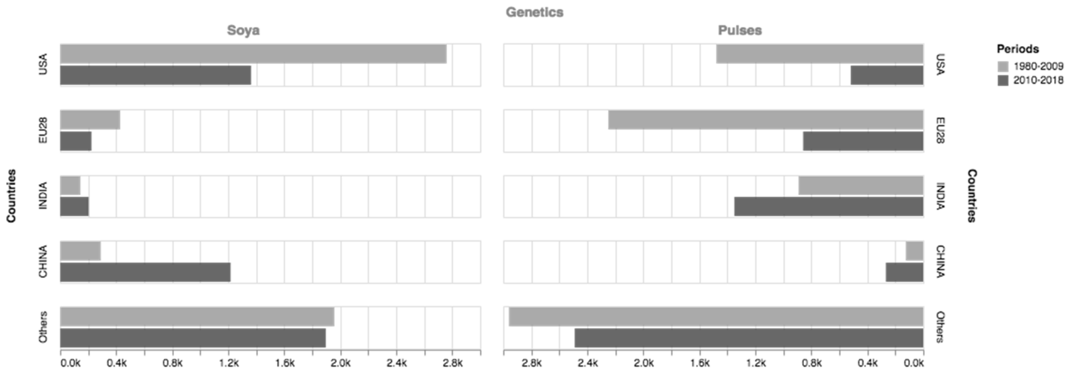
Productivity gains in agriculture remained very low until the agrarian revolutions of the 17th and 18th centuries in Europe. At this time, historians have found that legumes were second to cereals in consumption preferences, and in opposition to animal products. In France, for instance, paintings of the Renaissance period contrasted nobles who could go hunting with peasants reduced to eating lentils and bread [38]. It was a privilege of the nobility to consume meat more frequently, strongly linked to the right to hunt. This privilege has undoubtedly marked the collective unconscious towards a preference for the consumption of meat products, which in turn is also strongly correlated with the increase in incomes during the 20th century. Therefore, the current preference of Western countries for animal-based proteins is not only related to nutritional interest. However, consequently, high consumption of animal products makes it unnecessary to associate cereals and legumes consumption to meet protein needs.

In addition, during the succession of wars affecting Europe in the 19th and 20th centuries, legumes were frequently presented as filling foods during food shortages. Combined with a traditional image of "poor man's meat" or as a food related to famines and wars, after the Second World War consumers gave up legumes, and their consumption fell in Western countries. Trade agreements between Europe and the USA resulted in no development of soya in Europe during several decades, and thus to important soya imports for livestock.

Nowadays, Europe presents the lowest consumption with 3 kg/year per capita. In some European countries, consumption is even less, such as France (1.7 kg/year per capita in 2011, Agreste statistics). Globally, legumes are far more used for feeding animals, but with a minor position in feed formulas compared with soya. In addition to this trend, chemical fertilizers development favored a nitrogen cycle conception in cropping system without legumes (see [9] for more insights on economic trade-offs on legumes uses).

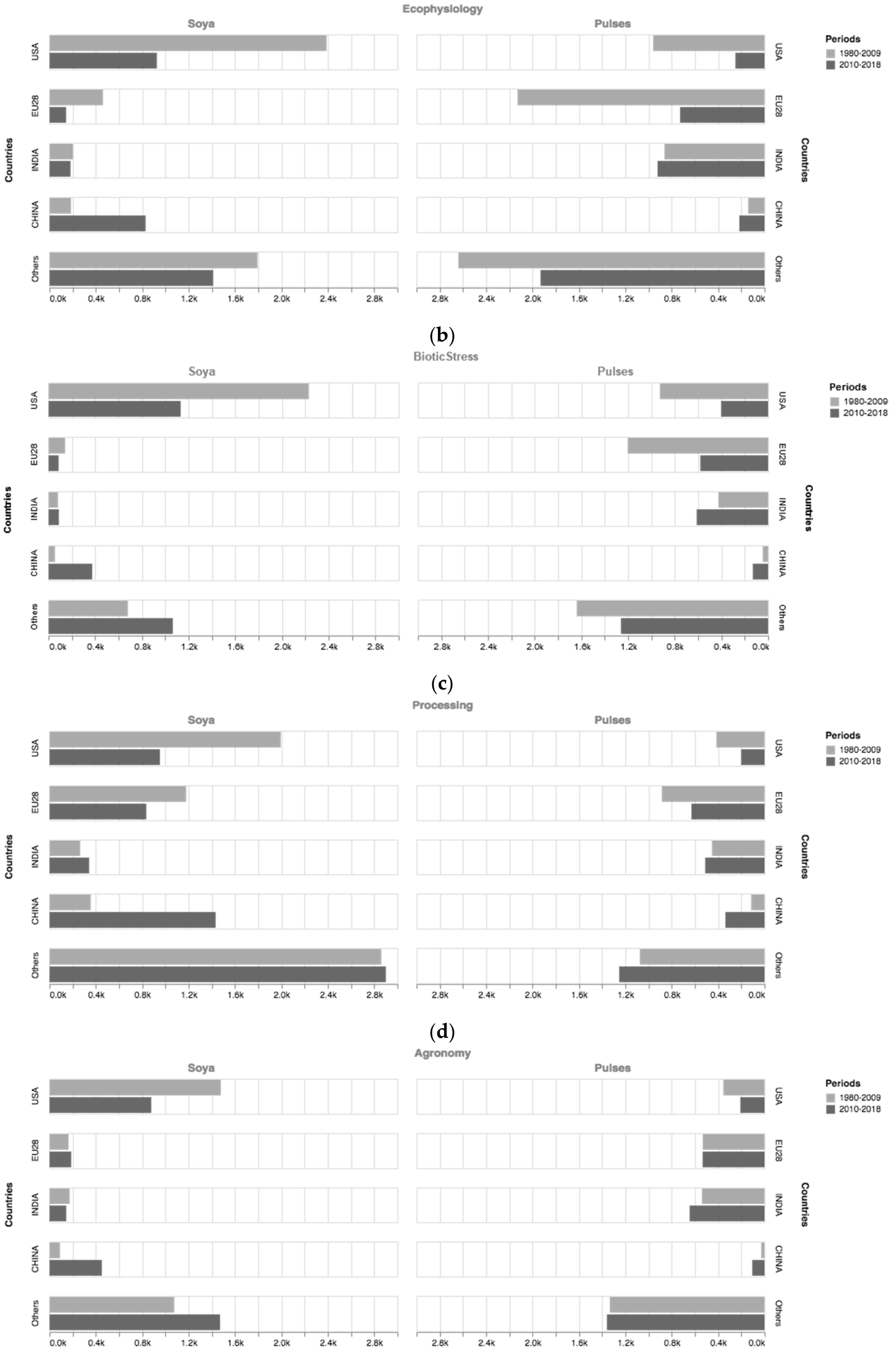
Appendix D. Corpus Broken Down by Theme and the Four Main Publishing Countries/Geographical Areas

Only records indexed with soya or the ones indexed with pulses were included (i.e., records co-indexed with several groups of grain-legumes were excluded); proportional count linked to international records applied; a group count done for the EU28; "Others" are all identified countries other than the USA, China, India, and those belonging to the EU28. Each theme counts for the number of records indexed with this theme (with or without co-indexing theme), representing the importance of the theme. Graphs are ordered by the amount of records by themes.



(a)

Figure A2. *Cont.*



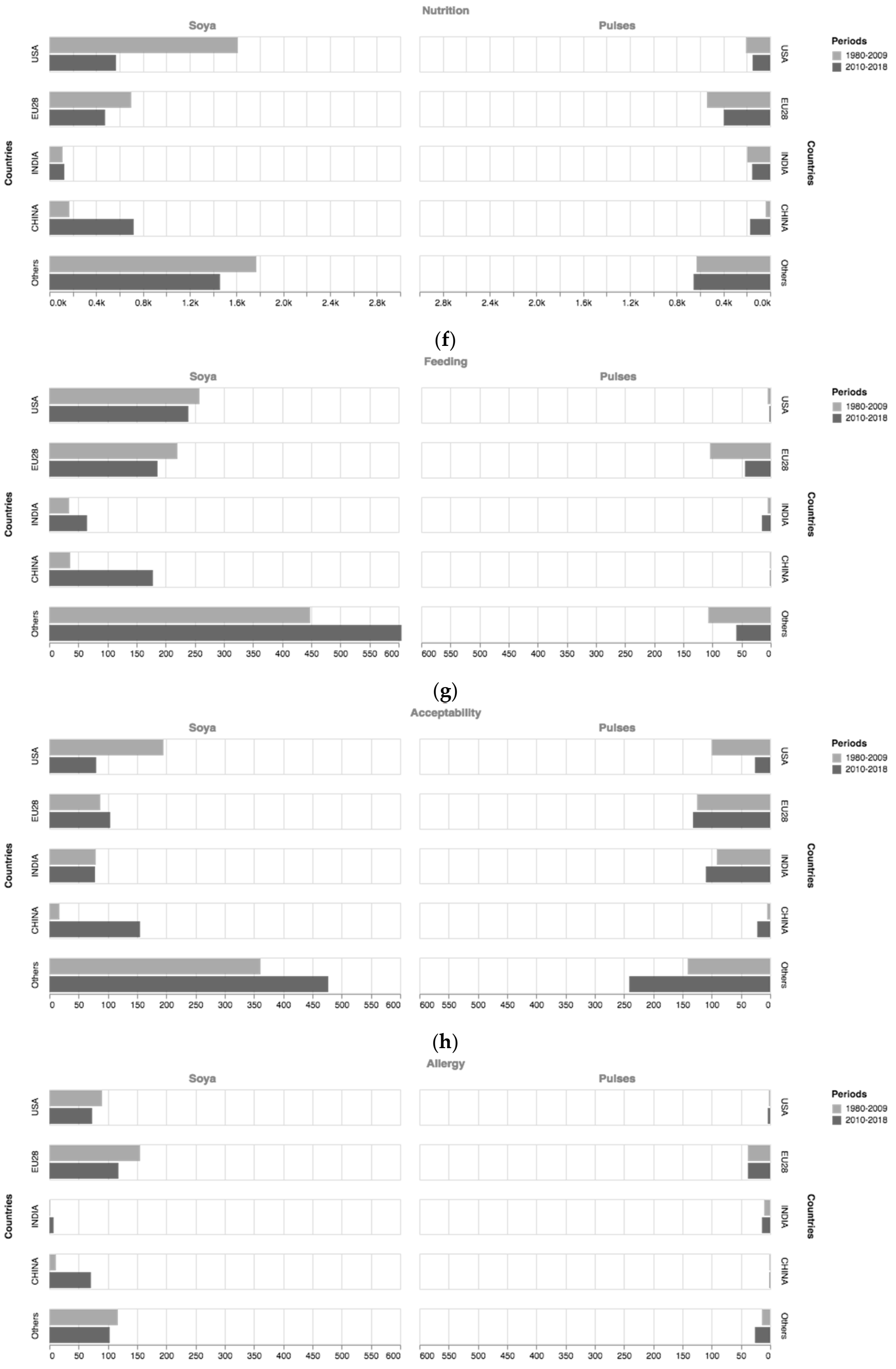
(**b**)

(**c**)

(**d**)

(**e**)

Figure A2. *Cont.*



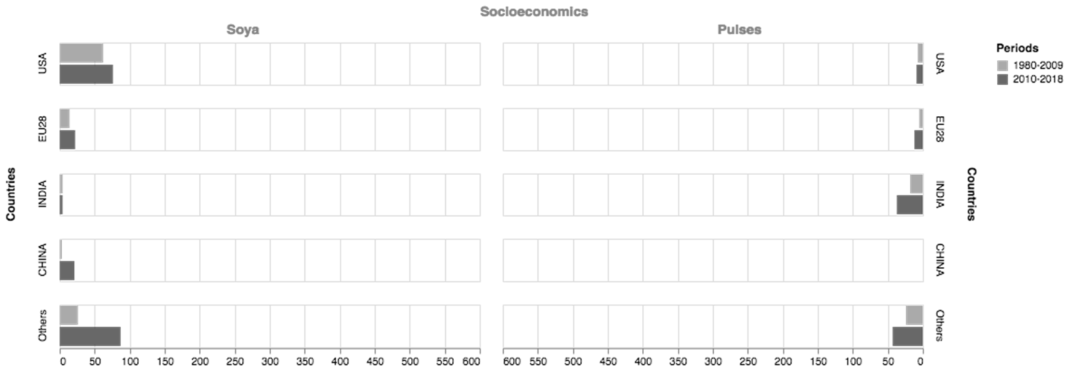
(**f**)

(**g**)

(**h**)

(**i**)

Figure A2. *Cont.*



(**j**)

Figure A2. Records number by theme and by period for the USA, the EU28, India, China, and Others. Scale based on 3 k for the figures (**a**–**f**), and on 0.6 k for the figures (**g**–**j**).

References

1. Lemaire, G.; de Faccio Carvalho, P.C.; Kronberg, S.; Recous, S. (Eds.) *Agroecosystem Diversity*; Academic Press: Cambridge, MA, USA, 2018.
2. Meynard, J.M.; Charrier, F.; Le Bail, M.; Magrini, M.B.; Charlier, A.; Messéan, A. Socio-technical lock-in hinders crop diversification in France. *Agron. Sustain. Dev.* **2018**, *38*, 54. [CrossRef]
3. Frison, E.A.; Cherfas, J.; Hodgkin, T. Agricultural biodiversity is essential for a sustainable improvement in food and nutrition security. *Sustainability* **2011**, *3*, 238–253. [CrossRef]
4. Peoples, M.B.; Hauggaard-Nielsen, H.; Huguenin-Elie, O.; Jensen, E.S.; Justes, E.; Williams, M. The Contributions of Legumes to Reducing the Environmental Risk of Agricultural Production. In *Agroecosystem Diversity*; Lemaire, G., de Faccio Carvalho, P.C., Kronberg, S., Recous, S., Eds.; Academic Press: Cambridge, MA, USA, 2018; pp. 123–143.
5. Rawal, V.; Bansal, V.; Thokchom, D. Biodiversity for Food and Agriculture and Food Security. An Exploration of Interrelationships, Background Study Paper NO. 69-2019, FAO Commission on Genetic Resources for Food and Agriculture. Available online: http://www.networkideas.org/wp-content/uploads/2019/02/ca3218en.pdf (accessed on 25 July 2019).
6. Weiner, J. Applying plant ecological knowledge to increase agricultural sustainability. *J. Ecol.* **2017**, *105*, 865–870. [CrossRef]
7. Magrini, M.-B.; Anton, M.; Chardigny, J.-M.; Duc, G.; Duru, M.; Jeuffroy, M.-H.; Meynard, J.-M.; Micard, V.; Walrand, S. Pulses for sustainability: Breaking agriculture and food sectors out of lock-in. *Front. Sustain. Food Syst.* **2018**, *2*, 64. [CrossRef]
8. Magrini, M.-B.; Befort, N.; Nieddu, M. Technological Lock-In and Pathways for Crop Diversification in the Bio-Economy. In *Agroecosystem Diversity*; Lemaire, G., de Faccio Carvalho, P.C., Kronberg, S., Recous, S., Eds.; Academic Press: Cambridge, MA, USA, 2019; pp. 375–388.
9. 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. Why are grain-legumes rarely present in cropping systems despite their environmental and nutritional benefits? Analyzing lock-in in the French agrifood system. *Ecol. Econ.* **2016**, *126*, 152–162. [CrossRef]
10. Rawal, V.; Navarro, D.K. *The Global Economy of Pulses*; FAO: Rome, Italy, 2017.
11. Watson, C.A.; Reckling, M.; Preissel, S.; Bachinger, J.; Bergkvist, G.; Kuhlman, T.; Lindström, K.; Nemecek, T.; Topp, C.F.; Vanhatalo, A.; et al. Grain legume production and use in European agricultural systems. *Adv. Agron.* **2017**, *144*, 235–303.
12. Dosi, G.; Nelson, R.R. Technical Change and Industrial Dynamics as Evolutionary Processes. In *Handbook of the Economics of Innovation*; Hall, B.H., Rosenberg, N., Eds.; Elsevier: Standford, CA, USA, 2010; Volume 1, pp. 51–127.
13. Rogers, E.M. *Diffusion of Innovations*; Free Press: New York, NY, USA, 2003.

14. Callon, M.; Law, J.; Rip, A. *Mapping the Dynamics of Science and Technology*; Palgrave Macmillan: London, UK, 1986.
15. Noyons, E.C.M. Science Maps within a Science Policy Context. In *Handbook of Quantitative Science and Technology Research, the use of Publication and Patent Statistics in Studies of S&T Systems*; Moed, H., Glänzel, W., Schmoch, U., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2004; pp. 237–255.
16. Noyons, E. Bibliometric mapping of science in a policy context. *Scientometrics* **2001**, *50*, 83–98. [CrossRef]
17. Rafols, I.; Hopkins, M.M.; Hoekman, J.; Siepel, J.; O'Hare, A.; Perianes-Rodríguez, A.; Nightingale, P. Big Pharma, little science: A bibliometric perspective on Big Pharma's R&D decline. *Technol. Forecast. Soc. Chang.* **2010**, *81*, 22–38.
18. Barbier, M.; Bompart, M.; Garandel-Batifol, V.; Mogoutov, A. Textual Analysis and Scientometric Mapping of the Dynamic Knowledge in and around the IFSA Community. In *Farming Systems Research into the 21st Century: The New Dynamic*; Springer: Cham, The Netherlands, 2012; pp. 73–94.
19. Glänzel, W.; Moed, H.F.; Schmoch, U.; Thelwall, M. *Springer Handbook of Science and Technology Indicators*; Springer: Basel, Switzerland, 2019. [CrossRef]
20. Zitt, M.; Lelu, A.; Cadot, M.; Cabanac, G. Bibliometric Delineation of Scientific Fields. In *Springer Handbook of Science and Technology Indicators*; Springer: Basel, Switzerland, 2019; pp. 25–68. [CrossRef]
21. Meyer, E.T.; Schroeder, R. Untangling the web of e-Research: Towards a sociology of online knowledge. *J. Informetr.* **2009**, *3*, 246–260. [CrossRef]
22. Wyatt, S.; Milojević, S.; Park, H.; Leydesdorff, L. Quantitative and Qualitative STS: The Intellectual and Practical Contributions of Scientometrics, SSRN 2015. Available online: https://ssrn.com/abstract=2588336 (accessed on 12 February 2019).
23. Cecere, G.; Martinelli, A. Drivers of knowledge accumulation in electronic waste management: An analysis of publication data. *Res. Policy* **2017**, *46*, 925–938. [CrossRef]
24. Minguillo, D.; Tijssen, R.; Thelwall, M. Do science parks promote research and technology? A scientometric analysis of the UK. *Scientometrics* **2015**, *102*, 701–725. [CrossRef]
25. Tancoigne, E.; Barbier, M.; Cointet, J.P.; Richard, G. The place of agricultural sciences in the literature on ecosystem services. *Ecosyst. Serv.* **2014**, *10*, 35–48. [CrossRef]
26. Martín-Martín, A.; Orduna-Malea, E.; López-Cózar, E.D. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: A multidisciplinary comparison. *Scientometrics* **2018**, *116*, 2175–2188. [CrossRef]
27. Johnson, R.; Watkinson, A.; Mabe, M. *The STM Report*, 5th ed.; An Overview of Scientific and Scholarly Publishing; International Association of Scientific, Technical and Medical Publishers: Hague, The Netherlands, 2018.
28. European Commission. Report from the Commission to the Council and the European Parliament on the Development of Plant Proteins in the European Union, COM 757. Brussels. Available online: https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/plants_and_plant_products/documents/report-plant-proteins-com2018-757-final_en.pdf (accessed on 5 December 2018).
29. Garfield, E.; Sher, I.H. KeyWords Plus™—Algorithmic derivative indexing. *J. Am. Soc. Inf. Sci.* **1993**, *44*, 298–299. [CrossRef]
30. Zohary, D.; Hopf, M.; Weiss, E. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*, 4th ed.; Oxford University Press: Oxford, UK, 2012.
31. Aykroyd, W.R.; Doughthy, J. *Les graines de légumineuses dans l'alimentation humaine*; Organisation des Nations-Unies pour l'Alimentation et l'agriculture: Rome, Italy, 2008.
32. Sonnino, A. Leguminose da Granella e Ricerca Agricola—Pulses and Agricultural Research. Atti del Seminario Leguminose da Granella—Sant'Angelo Lodigiano. 14 October 2016, pp. 45–50. Available online: https://sites.google.com/site/storiagricoltura/download-area/atti_seminari_mulsa (accessed on 5 July 2018).
33. Lascialfari, M.; Magrini, M.-B.; Triboulet, P. The drivers of product innovations in pulse-based foods: Insights from case studies in France, Italy and USA. *J. Innov. Econ. Manag.* **2019**, *1*, 111–143. [CrossRef]
34. Sabaté, J. (Ed.) Environmental Nutrition. In *Connecting Health and Nutrition with Environmentally Sustainable Diets*, 1st ed.; Academic Press: Cambridge, MA, USA, 2019.

35. Sivasankar, S.; Ellis, N.; Buruchara, R.; Henry, C.; Rubiales, D.; Sandhu, J.S.; Negra, C. 10-year Research Strategy for Food Crops. 2016. Available online: https://pulses.org/future-of-food/10-year-research-strategy (accessed on 9 August 2019).
36. Meynard, J.M.; Jeuffroy, M.H.; Le Bail, M.; Lefèvre, A.; Magrini, M.B.; Michon, C. Designing coupled innovations for the sustainability transition of agrifood systems. *Agric. Syst.* **2017**, *157*, 330–339. [CrossRef]
37. Ciarli, T.; Ràfols, I. The relation between research priorities and societal demands: The case of rice. *Res. Policy* **2019**, *48*, 949–967. [CrossRef]
38. Quellier, F. Petite et grande histoire des légumineuses en Occident, Journées de la Fondation Louis Bonduelle, Paris, 2016. Available online: http://www.fondation-louisbonduelle.org/wp-content/uploads/2016/10/florent-quellier-histoire-des-legumineuses-rencontres-fondation-louis-bonduelle-legumes-secs-proteines-vegetales.pdf (accessed on 29 July 2019).