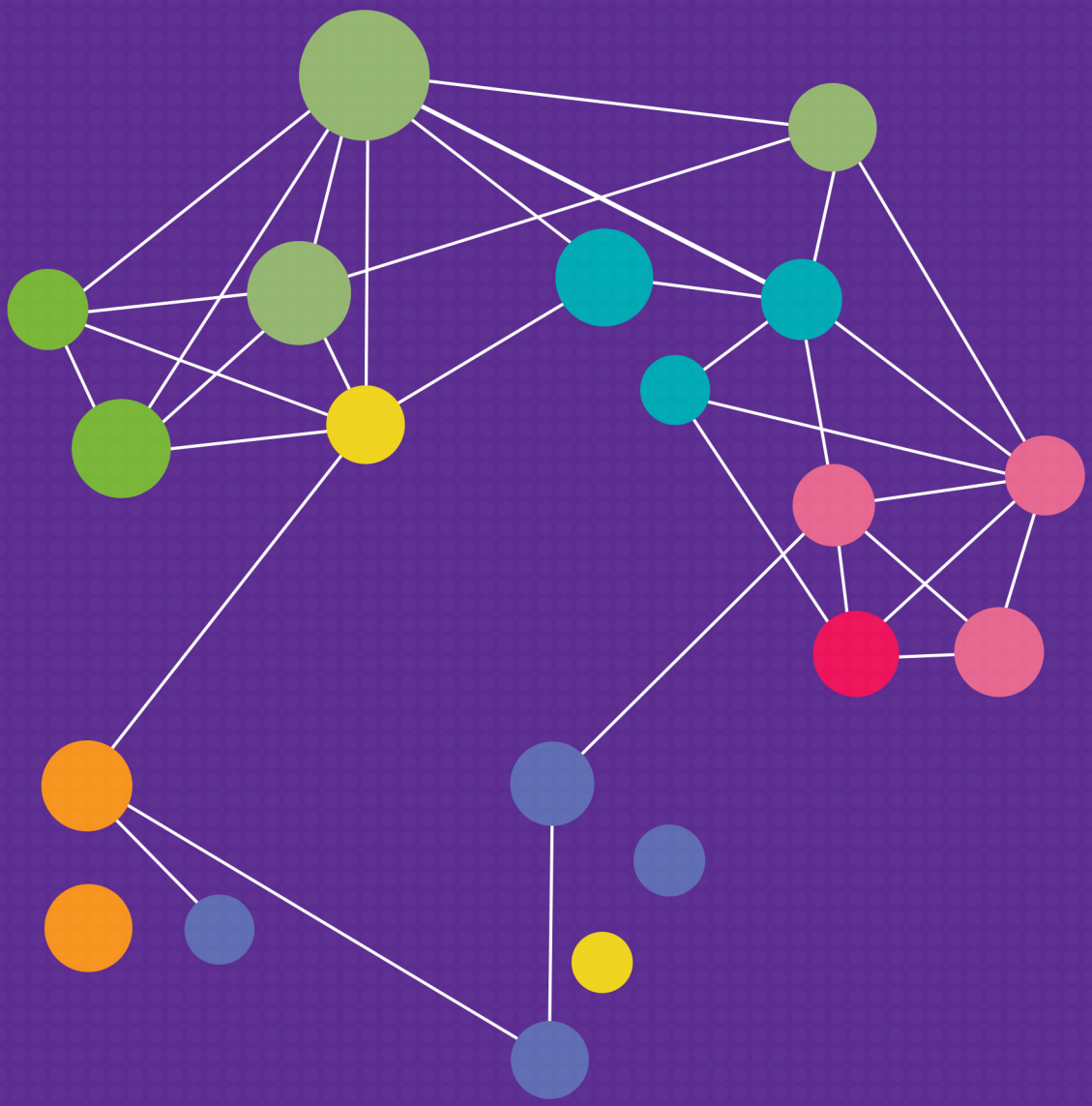
Observatoire des sciences et techniques

Caractériser les publications scientifiques relatives à un défi sociétal

Points OST, 2021.01



La collection Points OST est éditée par le Hcéres
Hcéres
2 rue Albert Einstein
75013 Paris

Auteur : Observatoire des Sciences et Techniques

Maquette de couverture : Éline Lazarini et Maguelonne Thiéry, Hcéres

ISBN : 978-2-492781-02-5

À citer comme suit :
OST (2021), Caractériser les publications scientifiques relatives à un défi sociétal, Points OST, 2021.01, Hcéres, Paris

Résumé

Pour évaluer des stratégies de recherche orientées par des défis sociétaux et pour mesurer leur impact sur la production scientifique, il est nécessaire d'identifier les corpus de publications scientifiques associées à ces choix de priorités de recherche. Cette étude présente une méthode pour constituer ces corpus qui sont transversaux aux disciplines scientifiques.

La méthode de délimitation de corpus développée à l'OST comporte plusieurs étapes. Une première étape sélectionne les publications à l'aide de mots-clés utilisés pour interroger une base de données. Une deuxième étape identifie les thèmes du corpus par l'utilisation d'un modèle probabiliste des fréquences des mots (*topic model* ou *modèle de thèmes révélés)*. Ensuite les éventuels thèmes hors du domaine sont repérés et les documents centrés sur ces thèmes sont retirés de la sélection initiale. Enfin une nouvelle application du modèle permet de proposer une analyse du domaine en le structurant en thèmes de recherche. La production de différents pays peut ensuite être caractérisée par leur spécialisation dans les thèmes identifiés.

La méthode est appliquée aux recherches sur les systèmes alimentaires qui apparaissent ainsi structurées en quatre thématiques : les systèmes alimentaires et leurs aspects politiques, sociaux et environnementaux ; les comportements des consommateurs ; les régimes et leurs impacts sur la santé ; les aliments et leurs procédés de transformation. Le positionnement des principaux pays producteurs dans le domaine révèle que l'Inde et la Chine sont spécialisées sur la thématique des aliments, que le Royaume Uni, l'Allemagne et les Pays-Bas le sont sur les aspects politiques, sociaux et environnementaux des systèmes alimentaires, alors que l'Espagne, la Suède et les États-Unis se concentrent sur les aspects santé des régimes alimentaires. La France a un profil semblable à la référence mondiale avec une légère spécialisation dans la thématique santé.

Mots-clés

Défi sociétal, systèmes alimentaires, délimitation de corpus de publications scientifiques, topic model.

A method to analyse scientific publications addressing societal challenges

Abstract

To assess research strategies guided by societal challenges and measure their impact on scientific production, it is necessary to build the corresponding corpus of scientific publications. This paper presents a method for delineating such a corpus.

The method developed at the OST for delineating the corpus consists of three main steps. The first selects keywords from a policy document and uses them as queries to retrieve documents from a publication database. The second one fits an LDA topic model to the corpus to identify its topics; irrelevant topics are then removed. During the third step, topic modelling is used again to cluster topics into main themes with respect to their semantic proximities and co-occurrence in documents. The corpus can then be used to compare the thematic profiles of countries.

The method is applied to research on food systems. Four main themes are distinguished within the food systems corpus: political, social and environmental aspects of food systems; consumer behaviour; diets and health risks; food chain efficiency. India and China are more interested in the subject of improving food chain efficiency. The UK, Germany and the Netherlands are specialised in the political, social and environmental aspects of food systems. Spain, Sweden and the USA publish more on healthy food. The profile of France is similar to that of the world in general, with a certain focus on health issues.

Keywords

Societal challenge, food systems, corpus delineation, topic model

Contributions et remerciements

Cette étude a été menée par Agénor Lahatte et Élisabeth de Turckheim. Isabelle Mézières a assuré le secrétariat de rédaction. Frédérique Sachwald a suivi la préparation et la rédaction en tant que directrice de l'OST.

La version finale a bénéficié des commentaires de trois relecteurs, Emmanuelle Janès-Ober et deux membres du Conseil d'orientation scientifique de l'OST, Nicole Haeffner-Cavaillon et Peter van den Besselaar. Les rapports des relecteurs et les réponses des auteurs ont fait l'objet d'une discussion lors d'une des réunions du Conseil d'orientation scientifique, ce qui a permis d'enrichir le texte de ce document.

La méthode de délimitation d'un corpus de publications avait été initialement élaborée dans le cadre d'une étude sur l'évaluation des choix stratégiques en matière de recherche menée pour le ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation. L'OST remercie ses correspondants au MESRI, Benoît Leblanc et Florent Olivier (SSRI) et les experts qui ont contribué aux ateliers pour le défi « Systèmes alimentaires et défi démographique » de la Stratégie nationale de recherche (SNR 2015), Michel Beckert et Isabelle Hippolyte. Aouatif de La Laurencie était cheffe du projet de l'OST sur la Stratégie nationale de recherche.

Sommaire

Introduction

Au cours de la décennie 2000, les politiques publiques ont développé différents instruments pour stimuler l'excellence scientifique et ses mesures. Par exemple, après plusieurs années de gestation au début des années 2000, le Conseil Européen de la Recherche a été créé en 2007 dans le cadre du septième programme-cadre[1] pour soutenir une « recherche à la frontière de la connaissance ». Au sein du programme suivant, Horizon 2020 (H2020), le Conseil Européen de la Recherche s'est inscrit au sein du premier pilier qui vise l'excellence scientifique et son budget a été sensiblement augmenté.

Sans se départir de l'objectif d'excellence, depuis la fin de la décennie 2000, les politiques publiques en faveur de la recherche ont progressivement mis l'accent sur des objectifs de résolution des défis sociétaux et des défis globaux, voire sur des missions consistant à répondre à des problématiques plus appliquées. Cette orientation peut être constatée dans de nombreux pays et à l'échelle de l'Union Européenne, par exemple à travers le contenu et l'organisation des programmes-cadres en faveur de la recherche et de l'innovation. Cette orientation se double en outre d'un soutien renforcé au transfert de technologie et à la création d'entreprises issues de résultats de recherche. L'évolution des instruments de politique de la recherche et de l'innovation en France et au niveau européen depuis une décennie illustre ces orientations des politiques publiques.

Dans la première moitié des années 2010, le programme-cadre européen H2020 comme la Stratégie nationale de recherche France-Europe 2020[2] ont retenu de grands défis sociétaux pour définir des priorités de recherche. La conception d'un Conseil européen de l'innovation et la préparation du programme Horizon Europe prolongent les orientations stratégiques précédentes. En France, les Programmes Prioritaires de Recherche (PPR) lancés à partir de 2017 cherchent à répondre à des défis sociétaux comme le programme dédié à la lutte contre la résistance aux antibiotiques.

Les politiques publiques sont ainsi devenues progressivement plus explicites dans la formulation des orientations des programmes de financement de recherches scientifiques. Cette orientation a des implications importantes sur les modalités de mesure de l'impact des activités de recherche. La qualité de la recherche peut ainsi s'entendre non seulement en termes d'impact scientifique mais aussi en termes de réponse à des besoins socio-économiques, ce qui pourrait être mesuré par exemple par la « résonance » des résultats avec les préoccupations de certains secteurs de la société (Bornmann et Haunschild, 2017). Néanmoins, l'attention portée à de nouvelles mesures de la qualité de la recherche en termes d'impact socio-économique semble parfois négliger les interactions fortes entre impact scientifique et potentiel d'impact en matière d'innovation sociale ou économique (Jonkers et Sachwald, 2018).

En amont des instruments de mesure des impacts de la recherche, la formulation des politiques de recherche en termes d'objectifs sociétaux pose la question des champs de connaissances mobilisés. Pour évaluer les résultats d'une stratégie de recherche, il faut pouvoir identifier, mesurer et caractériser les résultats de recherche des différents acteurs (pays, institutions, laboratoires) sur les priorités affichées. Il s'agit donc de construire des corpus de résultats, notamment de publications ou d'inventions, associés à ces priorités. Ces corpus, définis par des objectifs socio-économiques, ne correspondent pas à un ensemble de disciplines ou de spécialités disciplinaires, mais sont constitués de résultats issus de disciplines scientifiques différentes ou d'approches interdisciplinaires.

1 https://fr.wikipedia.org/wiki/Septième_programme-cadre

2 https://www.enseignementsup-recherche.gouv.fr/cid86688/strategie-nationale-recherche-france-europe-2020.html

Cette contribution porte sur une méthode permettant de constituer des corpus de publications correspondant aux domaines de recherche relatifs à des priorités de politique publique telles que celles qui ont pu être formulées pour répondre à des défis sociétaux. Une première partie décrit la méthode développée à l'OST pour construire un corpus de publications associé à un défi global et l'applique aux recherches sur les systèmes alimentaires. La deuxième partie montre comment l'analyse textuelle avec un *topic model* ou *modèle de thèmes révélés* permet de décrire et cartographier le domaine selon ses principales thématiques. La troisième partie présente brièvement les profils des recherches des principaux pays producteurs dans le domaine. La discussion propose quelques approfondissements pour évaluer cette méthode et enrichir l'analyse et l'évaluation de recherches thématiques.

1. Construction d'un corpus de publications associé à un défi sociétal

La méthode développée à l'OST comporte une première étape de sélection de publications à l'aide de mots-clés utilisés pour interroger une base de données. Une deuxième étape identifie les thèmes (ou topics) traités dans ce premier corpus par une analyse textuelle. Enfin le corpus final est obtenu par élimination des publications associées aux thèmes hors du domaine.

Cette méthode a été présentée lors de la conférence ISSI 2019, co-organisée par l'*International Society on Scientometrics and Informetrics* et le *European Network of Indicator Designers* (Lahatte, Turckheim et Chalumeau, 2019). Elle est illustrée par la construction d'un corpus associé à un axe de la Stratégie nationale de recherche française intitulé « Concevoir des systèmes alimentaires plus sains, durables, innovants et moteurs du développement économique » (SNR, 2014).

La construction du corpus s'appuie sur un document stratégique qui précise les enjeux de recherche. Pour l'axe relatif aux systèmes alimentaires du défi « Systèmes alimentaires et défi démographique » (MENESR, 2014), les questions posées à la recherche étaient les suivantes : Quels sont les systèmes alimentaires durables qui peuvent assurer la sécurité alimentaire et nutritionnelle des populations ? Quelles politiques et quelles actions institutionnelles les favorisent ? Quels sont les impacts sociaux et environnementaux des différents systèmes alimentaires ? Comment améliorer l'efficacité de la chaîne alimentaire, de la production et de la distribution jusqu'au traitement des déchets ? Comment et pourquoi les régimes alimentaires changent-ils ? Quels sont les déterminants des comportements des consommateurs ? Quels sont les impacts des régimes sur la santé et l'état nutritionnel des consommateurs ?

Pour rechercher les publications répondant à ces questions, l'OST a interrogé sa base de publications[3] avec une liste de termes choisis avec des experts du Ministère de l'enseignement supérieur, de la recherche et de l'innovation (MESRI), puis utilisés pour formuler des requêtes. Les termes sont recherchés dans le texte obtenu en fusionnant les métadonnées suivantes d'une publication : le titre, le résumé et les mots-clés des auteurs. La sélection de mots clés doit être large pour explorer la diversité des approches possibles. Ainsi, les requêtes sont des termes du vocabulaire général plutôt que des mots-clés spécifiques de disciplines scientifiques. Le tableau 1 présente les requêtes retenues pour construire un corpus sur les *Systèmes alimentaires durables.* Elles ont été proposées par l'OST et le jeu complet a été revu et validé par les experts du MESRI. Ces requêtes génèrent un corpus de 24 980 publications pour la période 2012-2017.

Du fait de la relative généralité de ces termes, les requêtes collectent aussi des publications hors du domaine concerné. Par exemple, le multi-terme *diet pattern%* est utilisé dans des publications sur

3 Version enrichie du Web of Science de Clarivate Analytics https://www.hceres.fr/fr/souces-et-donnees

l'alimentation humaine, mais aussi en écologie ou en biologie des populations. Une étape de nettoyage est nécessaire pour améliorer la précision du corpus.

Tableau 1 : requêtes utilisées pour construire le corpus initial de la thématique Systèmes alimentaires

Questions du document d'orientation	N° Requête	Requêtes
Systèmes alimentaires durables, sécurité alimentaire et nutritionnelle	1	sustainable AND food system%
	2	sustainable food
	3	food system% AND (safe OR safety)
	4	food system% AND (secure OR security)
	5	food security OR nutrition security
Efficacité de la chaine alimentaire	6	(efficien% OR optim%) AND (food process% OR food production)
	7	eco\-design AND food
	8	food preservation AND (technolog% OR process%)
	9	preservation AND (food process% OR food technol%)
	10	optim% AND (food circuit% OR food supply)
	11	short AND (food circuit% OR food supply)
	12	food waste%
	13	food loss%
	14	local% (agrifood OR agri\-food OR agrofood OR agro\-food)
Comportement des consommateurs, déterminants de ces comportements	15	food availab% and (choice% OR consumer%)
	16	food supply AND (choice% OR consumer%)
	17	food (information OR advertising OR marketing OR promotion) AND (choice% OR consumer%)
	18	food quality AND (choice% OR consumer%)
	19	(food safety OR safe food) AND (choice% OR consumer%)
	20	(healthy food OR healthy eating) AND (choice% OR consumer%)
	21	consumer% behavio% AND food
	22	food choice% AND consumer%
Régimes et transitions alimentaires	23	(food OR diet OR dietary) pattern%
	24	(food OR diet OR dietary OR nutrition) transition%
Impacts sociaux et environnementaux des systèmes alimentaires	25	social impact% AND (diet OR dietary OR food system%)
	26	environment% impact% AND (diet OR dietary OR food system%)
	27	impact% on environment AND (diet OR dietary OR food system%)
Impacts sur l'état nutritionnel et la santé	28	health benefit% AND (diet OR dietary OR food pattern%)
	29	healthy (diet OR dietary OR food pattern%)
	30	health risk% AND (diet OR dietary OR food pattern%)
Politique et gouvernance	31	(food OR nutrition) polic%
	32	(food OR nutrition) governance

Syntaxe d'une requête : la requête sustainable AND food system% *ramènera tous les documents qui contiennent à la fois le terme* sustainable *et, n'importe où ailleurs dans le texte, le terme composé* food system, *où le terme* system *peut avoir toutes les terminaisons possibles* (systems, systemic…)

L'OST utilise un modèle probabiliste, un *topic model* (Blei et al. 2003, Blei 2012), pour révéler des thèmes, ou *topics,* traités dans les publications du corpus. Parmi ces thèmes, il est alors possible d'identifier ceux qui sont hors du domaine d'intérêt. Une procédure par mots-clés spécifiques, appelés anti-requêtes, permet d'éliminer du corpus les publications qui traitent des thèmes hors domaine.

Modèles de thèmes révélés (*Topic models)*

Un modèle de thèmes révélés *(topic model* dans la littérature*)* suppose que les mots employés dans un ensemble de documents d'un corpus sont générés par des thèmes utilisant des vocabulaires spécifiques. Chaque document peut combiner plusieurs thèmes. Le modèle le plus simple* est défini par les fréquences des termes dans chaque thème et par les poids des thèmes dans chaque document.

Les thèmes peuvent être interprétés à partir des termes les plus fréquents de leur vocabulaire - ou ceux qui sont à la fois assez fréquents et assez spécifiques du thème - et avec les titres des documents dont ce thème est le sujet principal, par exemple s'il représente plus de 80 % des mots du document.

Le nombre de thèmes du modèle est déterminé en fonction du grain auquel on souhaite examiner le corpus. Un trop petit nombre de thèmes conduit à des thèmes difficiles à interpréter, un trop grand nombre peut isoler des thèmes artificiels comme ceux qui rassemblent du vocabulaire scientifique général. Quelques essais permettent d'ajuster le nombre de thèmes interprétables, adapté au type d'analyse qu'on souhaite faire.

Il faut noter que, contrairement aux méthodes habituelles de classification de documents, les thèmes ainsi construits ne définissent pas une partition des documents, car le modèle autorise au contraire les documents à combiner des thèmes. En général, un document traite majoritairement de 3 ou 4 thèmes et on peut, pour l'interprétation, se restreindre à ceux qui occupent par exemple au moins 20 % ou 25 % du document. Le nombre de documents qui combinent 2 thèmes donnent une information sur les connexions entre ces thèmes par la fréquence de leurs co-occurrences dans les publications du corpus.

Les thèmes ne constituent pas non plus une partition du vocabulaire du corpus mais ils utilisent a priori tous les termes de ce vocabulaire avec des fréquences différentes. Ainsi, un même terme peut être utilisé dans des contextes différents prenant en compte une éventuelle polysémie du terme. Ceci permet d'utiliser quasiment tous les mots du corpus, en dehors des mots de liaison ou de quelques mots parasites ou *stop words* qui sont supprimés dans une phase préalable à l'ajustement du modèle.

* *LDA ou Latent Dirichlet Allocation en relation avec les distributions a priori (de Dirichlet) des poids des thèmes dans un document et des termes du vocabulaire d'un thème.*

L'ajustement d'un modèle à 25 thèmes du corpus des 24 980 documents rapportés par le jeu de requêtes (Tableau 1) fait apparaître 22 thèmes couvrant les questions posées dans le document d'orientation (MENESR, 2014) et 3 thèmes hors du domaine des systèmes alimentaires. Ces derniers regroupent des recherches en écologie, en génétique des plantes et sur le métabolisme humain et l'endocrinologie. Cette classification des thèmes a été validée avec les experts du ministère.

Des *anti-requêtes* sont construites à partir de termes spécifiques aux thèmes hors domaine. Les publications qui les utilisent sont éliminées du corpus[4]. De plus, les publications dans des revues de spécialités disciplinaires trop techniques ou trop marginales par rapport à la question des systèmes alimentaires sont aussi éliminées (Tableau 2). Après ce nettoyage, on obtient un second corpus, plus précis, de 20 500 documents.

4 Cette méthode des anti-requêtes avait été proposée par Milanez et al. (2016). Ces anti-requêtes sont facilement identifiées avec le logiciel LDAvis (Sievert & Shirley, 2014).

Tableau 2 : anti-requêtes utilisées pour éliminer les documents du premier corpus qui les contiennent

Thèmes hors domaine	Anti-requêtes	Spécialités disciplinaires supprimées
Écologie, biologie marine et biologie aquatique, evo-devo	prey%, predator%, trophic, foraging, benthic, coral, reef%, juvenile, zooplancton, phytoplancton, pelagic, invertebrate, neolithic, archeological, webs, nest%, catches	ECOLOGY MARINE & FRESHWATER BIOLOGY ZOOLOGY ENTOMOLOGY ORNITHOLOGY
Génétique et biologie des plantes	genome%, landraces, accession%, loci, allele%, allelic, qtl%, arabidopsis, transcriptom%, snps, nucleotide, chromosome%, microsatellite%, heterosis, heterozygo%, inbred, barcoding	PLANT SCIENCES GENETICS & HEREDITY
Métabolisme et endocrinologie	insulin, rats, mice, hdl (high density lipoprotein), ldl (low dentity lipoprotein), lipoprotein, nafld (non alcoholic fat liver disease), hfd (high fat diet), crp, (C reactive protein), adipose, adinopectin, leptin, tnf (tumor necrosis factor), homa (homeostasis model assessment), steatosis, interleukin, lps (lipopolysaccharide), pcos (polycystic ovary syndrome), wistar (a laboratory rat), ppar (peroxisome proliferator-activated receptor), cortisol, macrophage%	ENDOCRINOLOGY & METABOLISM

Une question délicate est celle du choix du nombre de thèmes du modèle [5]. Ce nombre de thèmes peut être choisi par optimisation d'un critère statistique associé à la vraisemblance du modèle (Zhao et al. 2015). Cependant un choix qui maximise un tel critère ne conduit pas toujours à des thèmes interprétables (Chang et al. 2009). Pour cette étude, des modèles de 20 à 30 thèmes conduisent à des thèmes qu'on peut associer aux questions posées dans le document de politique publique de référence. L'analyse tient aussi compte du rôle de ce paramètre dans la précision de la méthode de nettoyage puisque seuls seront visibles et pourront être éliminés des thèmes dont la taille est dans la gamme des thèmes identifiés par le modèle.

2. Cartographie des thématiques du domaine Systèmes alimentaires

L'ajustement d'un modèle de thèmes révélés sur le corpus *Systèmes alimentaires* obtenu permet de caractériser les questions abordées par le document d'orientation de la stratégie nationale (MENESR, 2014) et les relations entre ces questions. Le graphique 1 présente le résultat de l'ajustement de 20 thèmes par une carte où la position des thèmes signale les proximités de vocabulaire entre thèmes et où la taille des bulles est proportionnelle au poids des thèmes dans le corpus. Les segments entre les bulles et leur épaisseur signalent les paires de thèmes qui apparaissent le plus souvent dans des publications communes (liens de co-occurrences).

5 Nous remercions A. Bonaccorsi, membre du Conseil d'orientation scientifique de l'OST, de nous avoir recommandé de préciser ce point que nous n'avions pas abordé dans une version précédente du document.

Cinq grandes thématiques structurent le domaine. Une première thématique traite des systèmes alimentaires, de leur gouvernance, du rôle des marchés et des échanges commerciaux, des facteurs sociaux de l'insécurité alimentaire (thèmes 1, 2 et 3). On peut y rattacher le thème 11 qui traite des impacts environnementaux des systèmes alimentaires et le thème 18 de l'impact du changement climatique sur la production agricole pour son lien avec la sécurité alimentaire. Les contours des thèmes 1, 2, 3 et 11 sont bien alignés avec les premières questions du document d'orientation de la stratégie nationale de recherche. Le corpus sélectionné y ajoute la question de l'impact du changement climatique sur les ressources alimentaires car ce thème est souvent traité conjointement avec celui de la gestion des ressources et de la distribution des produits alimentaires ainsi qu'avec celui des impacts environnementaux. Cette première grande thématique représente environ un tiers du corpus (Tableau 3).

Une deuxième grande thématique répond à une des questions du document d'orientation. Elle concerne les consommateurs, les déterminants de leur comportement, en particulier ceux qui influencent l'alimentation des jeunes, les politiques d'éducation nutritionnelle et leurs effets. Quantitativement, cette thématique représente 15 % du corpus (Tableau 3).

Une troisième grande thématique traite des régimes et de leurs impacts sur la santé. Elle représente 18 % du corpus.

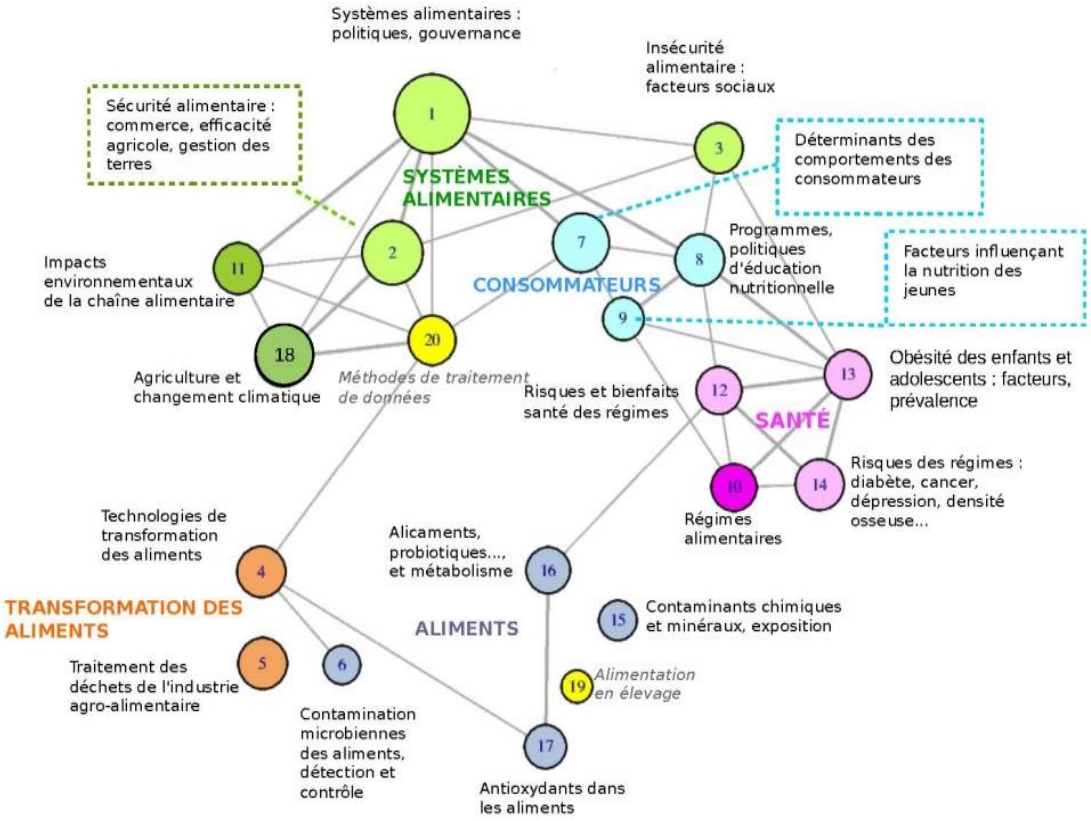
La quatrième grande thématique porte sur les aliments. Elle regroupe des travaux sur les aliments bénéfiques (antioxydants, alicaments), sur les contaminants chimiques et minéraux et sur les contaminations microbiennes des aliments et leur contrôle au cours des processus de transformation. Enfin, une thématique moins centrale regroupe deux thèmes concernant l'industrie agro-alimentaire : les technologies de transformation des aliments et la gestion des déchets de cette industrie. Le thème 6 sur les contaminations microbiennes des aliments industriels pourrait lui être rattaché. Il faut noter ici que l'on ne retrouve pas le thème sur l'efficacité de la chaîne alimentaire du document d'orientation alors que 9 requêtes ont été utilisées pour explorer la base de publications sur ce sujet. Si ces mots-clés, validés avec les experts, sont pertinents, l'étude suggère que cette question est peu abordée dans les publications scientifiques indexées dans la base. Avant de conclure que cette question est peu abordée par la recherche sur la période, il conviendrait d'explorer les publications dans d'autres sources de publications. L'analyse des financements de projets de recherche pourrait aussi permettre de compléter l'analyse sur ce point.

Dans ce corpus apparaît un thème portant sur les méthodes de traitement de données (thème 20) parce qu'il partage des documents avec plusieurs autres thèmes du corpus.

Enfin, un thème sur l'alimentation des animaux d'élevage traitant de questions de zootechnie (thème 19) a échappé au contrôle lors de la phase de nettoyage. Un grain plus fin, avec un modèle à 30 thèmes pour la première étape l'aurait identifié.

Les liens de co-occurrence de paires de thèmes représentés sur cette carte apportent une information sur leurs connexions. En particulier, les thèmes sur les aliments ou sur la chaîne de production qui ont des vocabulaires spécifiques sont peu connectés aux autres thèmes. Il n'apparaît pas de lien entre les thèmes sur les contaminations des aliments et ceux sur les régimes et la santé des consommateurs. Au contraire, les questions de comportement des consommateurs sont bien connectées aux enjeux de santé d'une part, aux enjeux de gouvernance et de politiques alimentaires d'autre part.

Graphique 1 : carte des thèmes du corpus Systèmes alimentaires



Source : base OST, Web of Science, calculs OST

Lecture : la taille des bulles est proportionnelle au poids des thèmes dans le corpus ; les couleurs indiquent le groupement en 5 thématiques de thèmes qui sont proches par leur vocabulaire. Sur les 20 thèmes, 2 (en jaune) ne seront pas utilisés dans la suite de l'analyse.

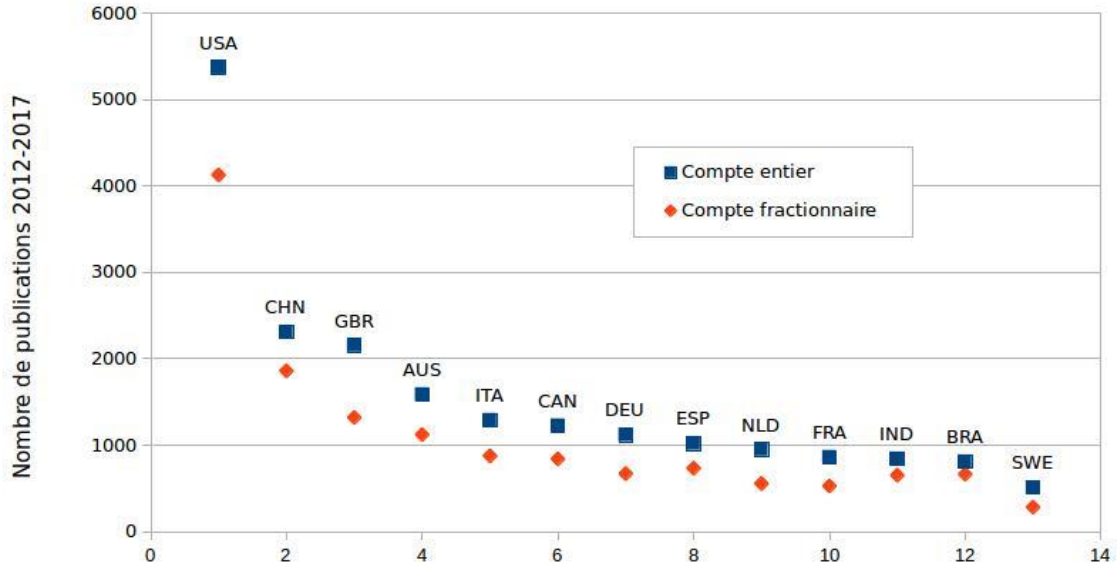
Ces relations de proximité peuvent être utiles pour explorer les connexions entre des thématiques particulières. Elles peuvent permettre de définir des priorités de recherche pour développer des connexions trop faibles et compléter la réponse de la recherche à un défi sociétal. C'est par exemple ce qu'a proposé Inrae dans une réflexion sur le Nexus Santé-Agriculture-Environnement[6]. Un document de recherche prospective souligne « l'intérêt qu'il y a à considérer le système alimentaire à l'aune des enjeux de santé globale » et identifie alors de grands enjeux de recherche pour connecter les nombreux travaux de recherche existants sur chacune des composantes de ce nexus (INRAE, 2020).

6 « La notion de Nexus est apparue récemment dans un certain nombre d'instances internationales en prolongement des publications du Millenium [...] pour tenir compte du fait que parmi les 17 objectifs mis en avant, certains sont étroitement liés, de façon positive ou négative, et doivent donc être associés dans la construction des priorités de recherche et la définition ou l'évaluation des politiques publiques »

3. Profils nationaux dans le corpus des publications sur les Systèmes alimentaires

Au sein du corpus *Systèmes alimentaires,* 13 pays ont publié plus de 500 articles au cours de la période 2012-17 (graphique 2). Les deux premiers producteurs sont, comme pour le total des publications, les États-Unis et la Chine, les premiers publiant plus que la seconde sur cette thématique. Le Japon, cinquième producteur mondial de publications en 2016 (OST 2019), est absent. L'Australie, le Canada, l'Espagne et les Pays-Bas publient plus que la France dans le domaine, alors que leur production totale de publications scientifiques était moindre sur la période[7]. L'Italie publie sensiblement plus que la France sur les systèmes alimentaires. L'ordre des pays, associé au nombre de publications dans le corpus Systèmes alimentaires, diffère ainsi de celui associé à la production scientifique globale des pays car certains pays comme l'Australie, les Pays-Bas et le Brésil sont plus mobilisés que d'autres - comme le Japon ou la Chine - sur le domaine.

Graphique 2 : nombre de publications des 13 premiers producteurs, 2012-17, selon le type de compte



Source : base OST, Web of Science, calculs OST

Lecture : une co-publication entre plusieurs pays est comptée 1 pour chaque pays avec le compte entier ; une fraction de la publication est attribuée à chaque pays avec le compte fractionnaire.

Les thèmes issus de l'ajustement d'un modèle de thèmes révélés sont aussi utiles pour comparer le positionnement des pays au sein du corpus global. Les contributions des pays peuvent être comparées par groupe de thèmes de façon à fournir une analyse plus synthétique.

Un tel regroupement permet de mieux contrôler la variabilité des thèmes due à la part d'aléatoire des algorithmes[8]. Le tableau 3 reprend les thématiques précédentes en groupant les thématiques sur les aliments et sur les processus de production, ce qui conduit à quatre groupes.

7 Le rang des pays pour le total des publications toutes disciplines confondues peut être consulté par exemple dans le rapport OST (2018), La position scientifique de la France dans le monde 2000-2015. Paris : Hcéres. https://www.hceres.fr/sites/default/files/media/downloads/Hc%C3%A9res_OST_Position_Scientifique_France__0.pdf

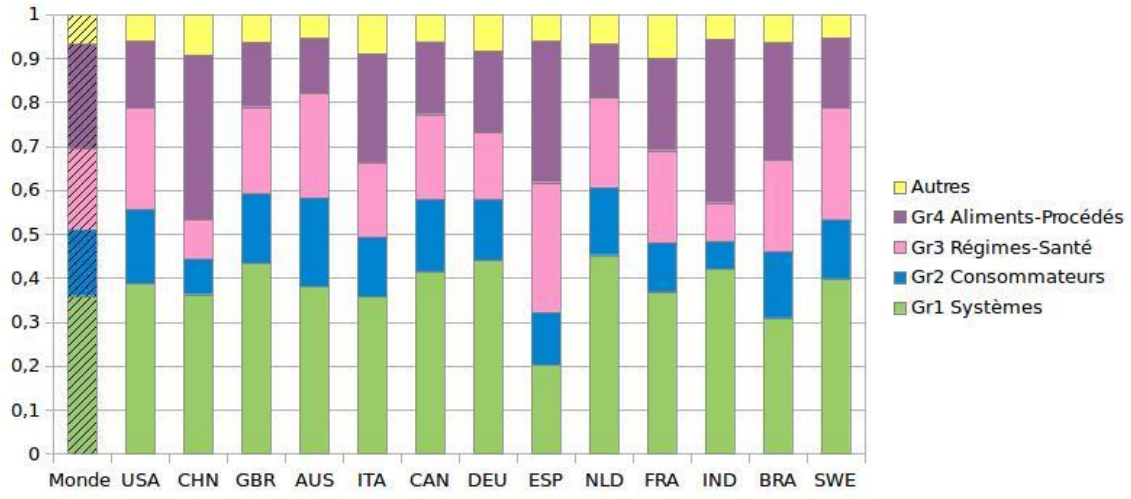
8 Dans les simulations permettant d'ajuster le modèle, chaque valeur initiale (*seed*) conduit à des thèmes différents qui peuvent reconfigurer la carte entre thèmes proches. Pour tenir compte de cette variabilité, il est utile de faire

Tableau 3 : quatre groupes thématiques au sein du corpus Systèmes alimentaires

	Thèmes	Groupe thématique	Part du groupe dans le corpus
Groupe 1	1, 2, 3, 11, 18	Systèmes alimentaires : politiques, gouvernance, impacts sociaux et environnementaux	36%
Groupe 2	7, 8, 9	Comportements des consommateurs	15%
Groupe 3	10, 11, 12, 13	Régimes et leurs impacts sur la santé	18%
Groupe 4	4, 5, 6, 15, 16, 17	Aliments et procédés de transformation des aliments	24%
Hors groupe	19, 20		7%

Les poids des groupes thématiques pour chacun des principaux producteurs définissent des profils pour chaque pays et pour le monde (graphique 3). Ainsi, la France a-t-elle une part relativement faible du corpus consacrée aux comportements des consommateurs, contrairement à l'Australie. L'Inde, la Chine et le Brésil ont une part importante de leur corpus consacrée aux aliments et procédés de transformation.

Graphique 3 : profil thématique des principaux pays producteurs



Source : base OST, Web of Science, calculs OST

Lecture : la production totale de chaque pays est ramenée à 1.

L'indice de spécialisation permet de comparer plus directement le profil d'un pays avec le profil mondial : il divise la part de ses publications relatives à un groupe dans le total des publications par ce même ratio pour le monde (graphique 4). Le Royaume-Uni, l'Allemagne et les Pays-Bas

plusieurs *runs* (avec des valeurs initiales différentes) et de construire les correspondances entre les thèmes des différents *runs*. Sur cette étude, nous avons vérifié que l'affectation des documents aux 4 thématiques est stable c'est-à-dire qu'elle a une faible variabilité.

apparaissent plus spécialisés que les autres pays dans les publications sur la thématique Systèmes, c'est-à-dire sur les politiques alimentaires et leurs impacts sociaux et environnementaux. Les pays les plus spécialisés sur les thématiques de santé sont les États-Unis, l'Australie, l'Espagne et la Suède. La Chine et l'Inde publient relativement peu dans les thématiques relatives à la santé et aux consommateurs. Ces deux pays sont à l'inverse les plus spécialisés dans la thématique relative aux aliments, leurs composants bénéfiques ou néfastes et les procédés de transformation alimentaire.

La France a un profil assez équilibré entre les thèmes. C'est sur le thème santé qu'elle est la plus spécialisée, mais dans une moindre mesure que d'autres grands producteurs (graphique 4). Seuls l'Italie et le Brésil ont des profils aussi équilibrés entre les thèmes.

Graphique 4 : indice de spécialisation des pays dans les groupes thématiques

Source : base OST, Web of Science, calculs OST

Lecture : la part de chaque pays dans une thématique est divisée par la part mondiale dans la même thématique.

Conclusion et approfondissements

Les modèles de thèmes révélés (topics models) constituent un outil intéressant pour construire et pour analyser des corpus associés aux recherches sur des défis sociétaux ou globaux. Ils permettent de prendre en compte la diversité des approches disciplinaires et interdisciplinaires de ces recherches et de décrire les connexions entre les thèmes. Cette propriété de ces modèles peut révéler une vision plus intégrée des problèmes et de leur étude. L'application de ces modèles à l'OST aide à la construction de corpus thématiques associés à des études diverses. Cela a notamment été le cas pour des études portant sur les recherches sur l'obésité (Cassi *et al.*, 2018) et sur le microbiote (Maddi, Sapinho & Baudoin, 2019). L'approche peut être adaptée à des domaines de recherche variés, ayant un périmètre précis ou au contraire très large. Une application de la méthode n'a pas permis de délimiter un corpus pertinent de publications scientifiques relevant de l'intelligence artificielle, du fait de la grande transversalité de ce domaine. En revanche, les thèmes révélés ont permis dans ce cas de souligner les différents domaines scientifiques concernés par les techniques de l'intelligence artificielle et leurs applications.

Plusieurs points de méthode n'ont pas été développés dans ce court texte qui fournit un exemple d'application de cette méthode de constitution de corpus de publications. Certains points soulevés par les relecteurs peuvent donner lieu à des approfondissements.

Une première remarque concerne la qualité de la procédure de nettoyage. Sa précision a été testée en évaluant pour chaque anti-requête, les proportions de documents retirés à tort ou conservés à tort dans le corpus. Les anti-requêtes retenues correspondaient à un compromis acceptable entre les deux mesures d'erreur. Une alternative intéressante proposée par une relectrice pour simplifier cette évaluation est de réaliser un sondage sur les documents les plus cités dont l'appartenance ou non au domaine serait définie à dire d'expert.

Un autre point concerne la qualité de *rappel* de la procédure, autrement dit, le fait que la méthode identifie bien l'essentiel des publications traitant du défi sociétal qui intéresse la politique publique. La vérification de la présence dans le corpus d'une liste de documents proposée par les experts consultés dans le cadre de l'exercice est recommandée. Quelques requêtes ont été ajustées avec une telle liste. Mais au-delà d'une telle vérification, se pose une question plus épistémologique : faut-il intégrer dans le corpus des travaux en amont sur des questions plus fondamentales dont peuvent dépendre des travaux de recherche finalisée ? Pour satisfaire une telle exigence, il serait possible d'enrichir le corpus des documents ciblés sur la finalité sociétale par la recherche des publications citées par les premières. Cela introduirait probablement des thèmes plus généraux comme le thème 20 sur les méthodes de traitement des données.

Symétriquement, afin de mieux inclure des travaux finalisés, il pourrait être pertinent d'explorer d'autres sources de données que les publications scientifiques indexées dans le WoS. Par exemple, dans le cas particulier des systèmes alimentaires et plus généralement des questions relevant de l'agriculture et de la santé, les publications du CAB (*CABI Global Health* et *CAB Abstracts)*[9] s'imposent, comme le soulignent Rafols et al. (2015).

Enfin, l'analyse des recherches sur les systèmes alimentaires s'est appuyée sur une carte de thèmes issus du modèle retenu. La méthode des thèmes révélés (*topic models*) a été préférée pour visualiser les thèmes de ce domaine car elle utilise tout le vocabulaire présent dans les métadonnées du corpus et car elle autorise différents thèmes à utiliser les mêmes termes, ce qui tient compte de la polysémie des mots. Elle identifie des thèmes dont les volumes sont plus homogènes que ceux obtenus avec les

9 Bases de données de l'organisation inter-gouvernementale (no-for-profit) *Commonwealth Agricultural Bureaux*

algorithmes de recherche de *communautés,* comme les algorithmes de Louvain (Blondel et al. 2008) ou de Leiden (Traag, Waltman & van Eck, 2019). Pour une étude approfondie du domaine des systèmes alimentaires, d'autres cartes pourraient néanmoins être mobilisées comme, par exemple, celles des réseaux scientifiques – réseaux de citations ou de copublications - ou encore celles qui positionnent ces recherches sur une carte générale de la science (*overlay map)*. En outre, certains logiciels permettent des analyses basées sur la fréquence des termes comme celles produites par les logiciels VosViewer [10] ou CorTexT [11]. Ces méthodes d'analyse, qui construisent des clusters de termes co-occurrents et des cartes de ces termes, sont proches de celle qui a été présentée mais peuvent être complémentaires.

10 https://www.vosviewer.com/
11 https://www.cortext.net/

Références

Axelos, M., Soler,L.-G., Dallongeville, Thomas, A., Akermann, G. et al. La santé, moteur des transitions agricole, alimentaire et environnementale.
Prospective scientifique interdisciplinaire. Décembre 2019. DOI 10.15454/fycc-jx29. https://hal.inrae.fr/hal-02864749

Blei, D. M., Nag, A. Y., & Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of machine learning research* 3, 993-1022

Blei D.M. (2012) Probabilistic Topic Models. *Communications of the ACM* **55** 77-83. doi.org/10.1145/2133806.2133826

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008) Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 10008, 6, https://doi.org/10.1088/1742-5468/2008/10/P10008

Bornmann, L. & Haunschild, R. (2017) Does evaluative scientometrics lose its main focus on scientific quality by the new orientation towards societal impact ?. *Scientometrics* 110: 937-43. doi.org/10.1007/s11192-016-2200-2

Cassi, L., Lahatte, A., Rafols, I., Sautier, P., de Turckheim, E. (2017) Improving fitness: Mapping research priorities against societal needs on obesity. *Journal of Informetrics* 1095-1213 https://arxiv.org/abs/1705.01151 ou doi.org/10.1016/j.joi.2017.09.010

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D. M. (2009) Tea Leaves; How Humans Interpret Topic Models. Advances in Neural Information Processing Systems 22 NIPS 2209 https://papers.nips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

Jonkers, K. & Sachwald, F. (2018) The dual impact of 'excellent' research on science and innovation: the case of Europe, *Science and Public Policy*, Volume 45, Issue 2, April 2018, Pages 159–174. doi.org/10.1093/scipol/scx071

Lahatte, A., de Turckheim, E., Chalumeau, L. (2019) Designing healthy and sustainable food systems: how is research contributing? In: *Proceedings of the 17th ISSI Conference*, Rome 2-5 September 2019, 523-534 https://drive.google.com/file/d/1nKOvCR14pJj2ayX33FXhU-uf7pmOoTE2/view https://prodinra.inra.fr/record/483530

Maddi, A., Sapinho, D., Baudoin, L. (2019) Mapping an emerging research subject: case of microbiota concept In: *Proceedings of the 17th ISSI Conference,* Rome 2-5 September 2019, 1232-1243 https://drive.google.com/file/d/1nKOvCR14pJj2ayX33FXhU-uf7pmOoTE2/view

MENESR (2014) Sécurité alimentaire et défi démographique, Bilan de l'atelier n° 5, Stratégie nationale de la recherche, https://cache.media.enseignementsup-recherche.gouv.fr/file/Strategie_Recherche/23/0/Rapport_atelier_5_314230.pdf

Milanez, D.H., Noyons, E. & de Faria, L.I.L. (2016) A delineating procedure to retrieve relevant publication data in research areas: the case of nanocellulose. *Scientometrics* 107: 627. doi:10.1007/s11192-016-1922-5

OST (2021) La position scientifique de la France dans le monde et en Europe, 2005-2018. Hcéres, Paris.

Rafols, I.; Ciarli, T & Chavarro, D. (2015) Under-Reporting Research Relevant to Local Needs in the Global South. Database Biases in the Representation of Knowledge on Rice." In International Conference on Scientometrics and Informetrics. Istanbul: International Conference on Scientometrics and Informetrics.

Sievert, C., & Shirley., K., E. (2014) LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* 63-70. Available at https://CRAN.R-project.org/package=LDAvis

SNR (2014) Stratégie Nationale de Recherche: bilan des 10 ateliers https://www.enseignementsup-recherche.gouv.fr/cid78802/www.enseignementsup-recherche.gouv.fr/cid78802/strategie-nationale-de-recherche-bilan-des-travaux-des-10-ateliers.htm

SNR (2015) Stratégie Nationale de Recherche - France Europe 2020 https://www.enseignementsup-recherche.gouv.fr/cid86688/www.enseignementsup-recherche.gouv.fr/cid86688/www.enseignementsup-recherche.gouv.fr/cid86688/strategie-nationale-de-recherche-france-europe-2020.html

Traag, V. A., Waltman L. & van Eck N. J. (2019) From Louvain to Leiden : guaranteeing well-connected communities. *Scientific Reports* 9:5233 https://doi.org/10.1038/s41598-019-41695-z

Zhao et al. (2015) A heuristic approach to determine an appropriate number of topics in topic modeling. *Bioinformatics* 16(Suppl 13):S8 http://www.biomedcentral.com/1471-2105/16/S13/S8

Haut Conseil de l'évaluation de la recherche et de l'enseignement suéprieur

Le Hcéres est l'autorité administrative indépendante chargée d'évaluer l'ensemble des structures de l'enseignement supérieur et de la recherche, ou de valider les procédures d'évaluations conduites par d'autres instances. Par ses analyses, ses évaluations et ses recommandations, il accompagne, conseille et soutient la démarche d'amélioration de la qualité de l'enseignement supérieur et de la recherche en France.

Le département OST produit des analyses et des indicateurs qui contribuent à la réflexion stratégique des acteurs de l'enseignement supérieur, de la recherche et de l'innovation, aux évaluations du Hcéres et à l'évaluation des politiques publiques.

2 rue Albert Einstein
75013 Paris, France
T. 33 (0)1 55 55 60 10

hceres.fr @Hceres_ Hcéres Hcéres