ETAT DE LA RECHERCHE SCIENTIFIQUE MAROCAINE ISSU DU WEB OF SCIENCE

Bernard DOUSSET, dousset@irit.fr

Institut de Recherche en Informatique de Toulouse, IRIT-SIG, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex 9 (France),

Mots clefs:
Intelligence économique, Veille scientifique et technique, Bibliométrie, Scientometrie, Analyse de réseaux, Indicateurs, Collaborations internationales, Co-auteurs, Productivité scientifique.
Keywords:
Competitive Intelligence, Science and technology watch, Bibliometry, Scientometry, Network analysis, Indicators, International collaborations, Co-authorsip, Scientific productivity.
Palabras clave:
Inteligencia Competitiva o "Económica", Vigilencia Científica y Tecnologica, Bibliometría, Cienciometría, Análisis de Redes, Indicators, Collaboratio Internationale, Co-autoría, Productividad científica.

Résumé :

Nous présentons ici un ensemble de méthodes de fouilles de textes qui sont largement utilisées dans notre laboratoire pour dresser un état des lieux de la recherche dans un contexte donné : domaine scientifique spécifique, zone géographique bien délimitée, équipe de recherche, transfert de technologie, ... Après avoir mené deux études proches, une première sur les collaborations internationales de l'INRA [J.-L. Multon, G. Lacombe, B. Dousset 2001 & 2002] puis une autre sur le Maroc [B. Dousset 2007] depuis la base de données PASCAL de l'INIST/CNRS, nous avons choisi de réitérer cette expérience en choisissant cette fois-ci d'analyser les WoS (Web of Science) de l'éditeur Thomson, afin d'illustrer ce que peuvent apporter nos nouveaux outils de text mining à l'intelligence économique et plus particulièrement à la veille scientifique. Cette base très utilisée pour évaluer la recherche via les cocitations et l'impact factor des revues, assure une bonne couverture des publications scientifiques. Une étude plus complète peut être réalisée de la même manière en intégrant des bases comme Science Direct, Google Scholar, Journal Citation Report (pour l'impact factor) ainsi que d'autres bases plus spécifiques comme Chemical Abstrat, Biosis, ... L'hétérogénéité de ces bases n'est pas vraiment un problème face aux puissantes fonctions d'analyse morphologique proposées par notre plateforme d'analyse bibliométrique « Tétralogie ». Malgré tout, le temps passé à la constitution d'un corpus multi-sources et à sa préparation grève inévitablement le budget et la durée d'une telle étude. Pour les mêmes raisons, nous nous sommes limités aux 8 dernières années soit de 2010 à octobre 2017 pour tout ce qui concerne l'étude en dynamique de l'évolution de la recherche. Bien qu'un peu biaisé (pro américain) le corpus étudié contient un nombre de publications largement suffisant pour dresser un état des lieux de la recherche et pour en établir son évolution récente. Nous nous sommes attaché à faire ressortir les collaborations internationales, les signaux forts et faibles, les nouveaux sujets de recherche, les journaux scientifiques les plus utilisés pour publier, l'évolution des équipes, … Cette liste n'est bien entendu pas exhaustive, nous avons aussi exploité, pour l'étude complète, le texte libre (titres et résumés) ainsi que des données plus technique comme le type de document, les éditeurs, les conférences, les villes (pour générer les régions marocaines et les états américains), les langues, les e-mails, les bailleurs de fonds, … Les différents indicateurs que nous proposons sont très évocateurs de l'état des lieux de la recherche Marocaine, une étude plus en profondeur (zooms à la demande) est toujours possible, mais pour pousser vraiment plus loin il convient de se focaliser sur chaque discipline afin de limiter la complexité essentiellement due à la multiplicité des acteurs (auteurs, laboratoires, journaux et congrès) et de la terminologie (thésaurus, mots clés, classifications, multi-termes, …).

1 INTRODUCTION

Le but de cette étude est d'illustrer certaines méthodes de fouille de texte utilisées en bibliométrie ainsi que les méthodes récentes de visualisation qui viennent d'être ajoutées à la plate-forme « Tétralogie ». En nous intéressant à la production scientifique marocaine, nous avions deux objectifs : montrer la diversité et l'efficacité des outils d'analyse disponibles à l'heure actuelle et discuter de certaines difficultés d'exploitation liées à la forme même des données sources. Nous avons constitué un corpus de publications scientifiques issu de la base bibliographique WoS de Thomson de 1980 à 2017 en faisant un zoom particulier sur une période récente de près de 8 ans (2010 à octobre 20017). Les documents retenus contiennent au moins une fois dans le champ adresse un organisme Marocain. Pour réaliser cette étude, nous avons largement utilisé notre plate-forme dédiée à la veille stratégique en essayant, chaque fois que possible, d'illustrer notre propos par des approches différentes de traitement des données et de visualisation des résultats. Nous ne pouvons présenter, ici, qu'une faible partie des nouvelles connaissances synthétiques obtenus dans cette étude, le but étant surtout d'informer sur l'ensemble des possibilités offertes en extraction de connaissances à partir des données textuelles de type « notices bibliographiques ».
Comme « Tétralogie » permet de se connecter à distance sur les éléments d'une telle analyse, les personnes intéressées peuvent, à loisir, venir consulter les résultats chiffrées déjà obtenus et éventuellement pousser plus loin certaines investigations en réalisant des zooms sur leurs centres d'intérêts particuliers (zone géographique, domaine de recherche, équipe, laboratoire, …).

2 CARACTERISTIQUES DU CORPUS ETUDIE

2.1 La Base Web of Science (Thomson)

Le Web of Science (WoS) est la principale base bibliographique utilisée dans l'évaluation de la recherche en se basant sur les réseaux de co-citations. Dernièrement elle a pas mal évolué en s'ouvrant vers d'autres bases (Medline, Journal Citation Report, …), ce qui a permis d'en étendre assez largement sa couverture scientifique. Sa partie centrale, le Core, est plus homogène au niveau du format et contient les co-citations de ses propres articles. Son inconvénient majeur est son orientation pro américaine qui induit un certain biais par rapport à d'autres bases plus objectives. Ses principales caractéristiques sont les suivantes :

- Plus de 10 000 revues indexées
- Disponible en ligne à l'Université de Toulouse ou sur Docadis
- Domaines couverts:
    - Médecine
    - Odontologie
    - Pharmacie
    - Chimie
    - Informatique
    - Mathématiques
    - Physique
    - Sciences de l'Ingénieur
    - Planète et de Univers
    - Sciences du Vivant
    - Sport
    - Gestion, Economie, Droit
    - Sciences de l'Information
    - …
- Champs de recherche:
    - Topic
    - Title
    - Author
    - Author Identifiers

    - Editor
    - Group Author
    - Publication Name
    - DOI
    - Year Published
    - **Address**

- Télédéchargeable par groupe de 50 notices
- Titre + résumé d'une dizaine de lignes
- Adresses de tous les laboratoires concernés
- Le champ AU: (auteurs) pose des problèmes de structure:
    - noms courts et longs
    - initiales avec ou sans point
    - liens adresses, liens auteurs, ...

2.2 Equation de recherche

- Les opérateurs disponibles sont les suivants : NEAR/x, SAME, NOT, AND, OR
- Rechercher "Morocco" dans le champ adresse
- Présence de variations morphologiques :
    - Morocco (50 192), Morroco (37), Moroco (3), Maroc (6 718)
- Limitation possible sur des thèmes de recherche
    - Via les mots clés
    - Via le texte libre (titre et résumé)
    - Via les classifications
    - Mais l'équation doit rester simple (peu de termes, d'opérateurs)
- Limitation dans le temps
    - Années (2016: 5 264), périodes, …
- Limitation à des sous bases
    - Core du WoS (47 517), MEDLINE (14 350), …
- Une bonne solution: tout récupérer puis filtrer off line.

2.3 Période choisie et volume obtenu

- Support et périodes télé-déchargées : WoS en ligne de janvier 1980 à octobre 2017.
- Format html avec tous les champs visualisés
- Taille du corpus : plus de 46 000 fiches bibliographiques
- Volume brut : plus de 8 Go
- Volume une fois reformaté : 165 Mo soit environ 2% de données utiles.

2.4 Format html d'une notice bibliographique

Submicron silica shell-magnetic core preparation and characterization

By: Bitar, A (Bitar, Ahmad)[1] ; Vega-Chacon, J (Vega-Chacon, Jaime)[1,2] ; Lgourna, Z (Lgourna, Zineb)[3] ; Fessi, H (Fessi, Hatem)[1] ; Jafelicci, M (Jafelicci, Miguel, Jr.)[2] ; Elaissari, A (Elaissari, Abdelhamid)[1]

COLLOIDS AND SURFACES A-PHYSICOCHEMICAL AND ENGINEERING ASPECTS
Volume: 537 **Pages:** 318-324
DOI: 10.1016/j.colsurfa.2017.10.034
Published: JAN 20 2018
View Journal Impact

Abstract

The magnetic extraction, purification, labeling and detection of biomolecules and cells are widely applied recently. In addition, the synthesis of silica magnetic particles for biomedical applications has been developed. In this paper, the synthesis of silica-coated magnetic nanoparticles was performed through three steps. First, an organic ferrofluid was prepared by the coprecipitation technique. Then it was used to prepare the magnetic emulsion. Finally, the hydrophobic magnetic nanoparticles were coated by silica shell using the sol-gel process. The prepared particles during the three steps were characterized in term of morphology, chemical composition and magnetic behavior under magnetic field. The results confirmed the formation of silica-coated magnetic nanoparticles, which are superparamagnetic even after silica encapsulation. The particles size was adjusted in order that a magnetic field can easily separate the particles at the same time they are stable colloids. Obtained particles can be used to extract and purified the nucleic acids, immobilized the enzymes, proteins and antibodies or to label the bacteria cells.

Keywords

Author Keywords: Silica shell; Magnetic nanoparticle; Silica; Nanoparticle
KeyWords Plus: IRON-OXIDE NANOPARTICLES; BIOMEDICAL APPLICATIONS; DRUG-DELIVERY; FERROFLUID EMULSIONS; TEMPERATURE; SEPARATION; FUNCTIONALIZATION; TOXICITY; LATEXES; DESIGN

Author Information

Reprint Address: Elaissari, A (reprint author)
+ Univ Lyon 1, CNRS, UMR 5007, LAGEP, F-69622 Lyon, France.
Addresses:
+ [1] Univ Lyon 1, CNRS, UMR 5007, LAGEP, F-69622 Lyon, France
+ [2] Sao Paulo State Univ UNESP, Inst Chem, Araraquara, SP, Brazil
[3] Labomine Lab, Lot 35, Zi Tassila 80000, Agadir, Morocco
E-mail Addresses: abdelhamid.elaissari@univ-lyon1.fr

Funding

Funding Agency	Grant Number
Coordenacao de aperfeicoamento de pessoal de nivel superior (CAPES)	88881.132878/2016-01

View funding text

Publisher
ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS

Categories / Classification
Research Areas: Chemistry
Web of Science Categories: Chemistry, Physical

Document Information
Document Type: Article
Language: English
Accession Number: WOS:000417068300037
ISSN: 0927-7757
eISSN: 1873-4359

Journal Information
Impact Factor: Journal Citation Reports

Other Information
IDS Number: FO7PJ
Cited References in Web of Science Core Collection: 35
Times Cited in Web of Science Core Collection: 0

Figure 1 : visualisation du format html des notices bibliographique du WoS

2.5 Interrogation du WoS

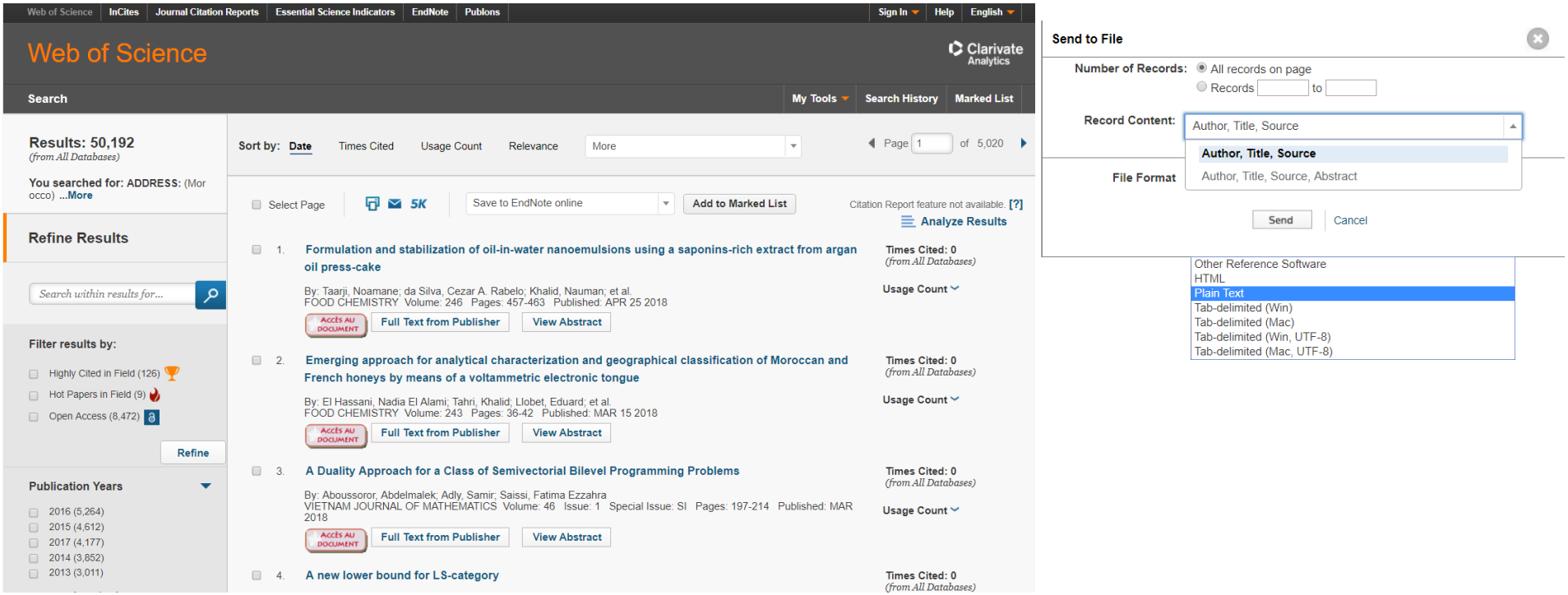


Figure 2 : mode de récupération standard des documents

2.6 Choix du format des notices

- Contenus par défaut:
  - Authors, Title, Source
  - Authors, Title, Source, Abstract
- Formats proposés:
  - Html
  - Plain text
  - Délimité par des tabulations (PC, Mac, utf8)
- Format html visualisé:
  - Toutes les adresses, E-mails de certains auteurs

- Noms des organismes de recherche, des éditeurs
- Plusieurs champs mots clés
- Impact Factor, DOI, ISSN, IDSN

2.7 Problème de reformatage de la base

Nous avons été conduits à reformater cette base pour les raisons suivantes :

- Le format html est très lourd et certaines informations inutiles se retrouvent dans chaque document
  - bandeaux, boutons, liens, explications, informations du site, …
- Comme le montre la figure 1, certaines balises sont absentes (titre, journal)
- D'autre sont peu informatives : by pour auteurs
- Certains noms d'auteurs sont des liens vers des pages suites
- Noms courts (initiale du prénom), noms long avec le prénom entier
- Ce même champ contient des liens vers les adresses
- D'autres informations se retrouvent dans des tableaux (Funding)
- Le champ adresse est éclaté et contient des liens
- L'impact factor n'est pas visible, mais il est bien présent dans le code source.

2.8 Format du champ adresse

- Il contient les adresses de tous les auteurs
- Plusieurs adresses possibles par notice (pour certaines, jusqu'à plusieurs centaines)
- Une balise "ADD: " par adresse
- Liens ente les auteurs et leurs adresses
- Exemples d'adresses:
  - ADD: TIM, ENSA, Marrakech, Morocco
  - ADD: IEMN DOAE, UVHC, F-59313 Valenciennes, France
  - ADD: LEOST, IFSTAR, F-59666 Villeneuve Dascq, France
- Les séparateurs sont les virgules et les chiffres
  - Les villes sont utilisées pour générer les régions marocaines et les états américains
  - Variations morphologiques sur les pays.

[1] Univ Adelaide, Dept Phys, Adelaide, SA, Australia
[2] SUNY Albany, Dept Phys, Albany, NY 12222 USA
[3] Univ Alberta, Dept Phys, Edmonton, AB, Canada
[4] Ankara Univ, Dept Phys, TR-06100 Ankara, Turkey
[5] Istanbul Aydin Univ, Istanbul, Turkey
[6] TOBB Univ Econ & Technol, Div Phys, Ankara, Turkey
[7] CNRS, IN2P3, LAPP, Annecy Le Vieux, France
……..
[276] Wigner Res Ctr Phys, Inst Nucl & Particle Phys, Budapest, Hungary
[277] Univ KwaZulu Natal, Discipline Phys, Durban, South Africa
[278] Univ Malaya, Dept Phys, Kuala Lumpur 59100, Malaysia

2.9 Méta données de Tétralogie pour le WoS

Sur la plate-forme Tétralogie, il est nécessaire d'établir un descripteur de la structure de la source utilisée (meta données) afin que les outils d'extraction puissent parfaitement s'adapter au traitement d'un format bibliographique particulier. Il faut en outre définir chaque champ (nom, séparateurs, utilité, filtrage) et trouver une balise de synchronisation pour délimiter chaque document. Voici le descripteur Tétralogie pour la base PASCAL (en gras les champs qui ont changé de nom).

TI :
descripteurs des champs du WoS via le format html

# nom	abrev	champ	visible	Separateurs #
Multi-termes	MT	MTM:	True	b”
Titre	TI	TI:	True	"
Auteurc	AC	AU:	True	;"("
Auteurl	AL	AU:	True	;"\n")"
Source	SO	SO:	True	"
Journa-SO	JN	SO:	True	"
Journal	JI	JI:	True	"
Volume	VO	VO:	False	"
Is	IS	IS:	False	"
Pg	PG	PG:	False	"
Doi	DO	DOI:	False	"
Date_Pub	DP	DP:	True	b"-"\n")":","
Per_Pub	PP	DP:	True	b"-"\n")":","
Resume	AB	AB:	True	b","
AN	AN	AN:	True	:"ORD2"
An	An	An:	True	"
Pi	PI	PI:	False	"
Doc_type	DT	DT:	False	"
Langue	LG	LG:	False	"
Mots_Cles	MC	KWa:	True	;","KWp:"MML:"\n"
Mots_AU	MA	KWa:	True	;","ADD:"Auth"RPADD:"
Mots_Plus	MP	KWp:	True	;","ADD:"Auth"RPADD:"
RP-add	RA	RA:	False	"
Editeur	ED	PU:	False	"
Adresse	AD	ADD:	True	]"
Organisme	OR	ADD:	True	,"."["\n"0"1"2"3"4"5"6"7"8"9"Orga")"]"/"
Pays	PA	ADD:	True	,"."["\n"0"1"2"3"4"5"6"7"8"9"Orga")"]"/"
USA	US	ADD:	True	b","
Bresil	BR	ADD:	True	b","
Chine	CN	ADD:	True	b","
Maroc	MA	ADD:	True	]",";"0"1"2"3"4"5"6"7"8"9"
Organisme2	or	OR:	True	["\n"]"
E-Mail	EM	MAIL:	True	;"b"
Funding	FU	FU:	False	;"b"
AU-ident	AI	AI:	False	;"b"
Fd	FD	FD:	False	"
Editeur	ED	PBR:	True	,"ORD1"
Cat	CA	CAT:	False	"
Rea	RE	REA:	False	"
ImpactF	IF	IF:	True	b"ORD1"
Issn	IS	ISSN:	False	"
St	ST	ST:	False	"
FIN	FIN	FIN	FIN	"

Le champ multi-termes MTM : a été ajouté, il correspond au résultat du traitement sémantique des champs en texte libre (titre et résumé) afin d'en extraire les mots composés (multi-termes) qui vont permettre une indexation à jour du corpus permettant de détecter l'innovation absente des champs d'indexation proposés par le WoS (Mots-clés KWa :, Mots-clés plus KWp :, , …).
Dans le descripteur de format ci-dessus, la balise True permet de travailler sur le champ associé, la balise False le masque dans tous les menus du logiciel. Pour les séparateurs, certains jockers sont utilisés : ORD_i permet d'extraire uniquement le $i^{ième}$ segment de texte du champ découpé suivant les séparateurs proposés (ORD_0 pour extraire le dernier segment), \n désigne le changement de ligne, \" le double guillemet, b le blanc, …

2.10 Répartition des articles dans le temps

Pour la période étudiée (janvier 1980 à octobre 2017) nous avons récupéré plus de 45 000 notices bibliographiques contenant Morroco ou équivalent dans le champ adresse. Ci-dessous, nous illustrons la répartition de ces documents dans le temps. L'histogramme obtenu ne doit pas être interprété sans prendre en compte le retard d'indexation inhérent à toute base bibliographique et dû au décalage entre la parution d'un article et son entrée dans la base. Pour le WoS, nous estimons ce retard à environ 2 mois mais inégalement répartis entre les publications de premier plan (délai moins long) et les autres. Le déficit constaté sur 2007 à 2 origines : le délai que nous venons de mentionner et le fait que l'année n'est pas complète (seuls les mois de janvier à octobre étant pris en compte).

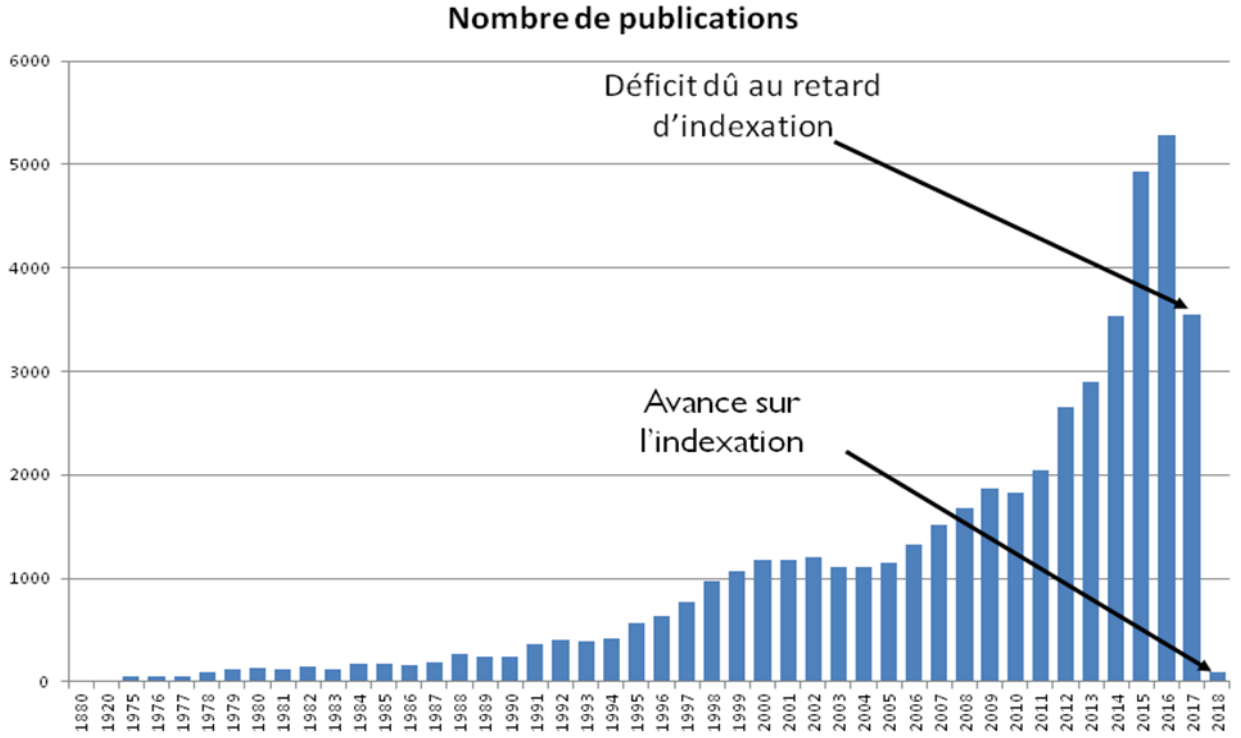


Figure 4 : évolution de la production scientifique marocaine

3 QUELQUES RESULTATS QUANTITATIFS

3.1 Le nombre de publications par journal

Bien que nous soyons dans une base bibliographique, le champ journal demande à être légèrement corrigé de quelques erreurs morphologiques, comme le montre le dictionnaire de synonymes suivant:

ACTA PHYSICA POLONICA A ACTA PHYSICA POLONICA
ACTA PHYSICA POLONICA B ACTA PHYSICA POLONICA
ANATOMIA HISTOLOGIA EMBRYOLOGIA-JOURNAL OF VETERINARY MEDICINE SERIES C-ZENTRALBLATT FUR VETERINARMEDIZIN REIHE C
ZENTRALBLATT FUR VETERINARMEDIZIN REIHE C-JOURNAL OF VETERINARY MEDICINE SERIES C-ANATOMIA HISTOLOGIA EMBRYOLOGIA
ANNALES DE CHIRURGIE PLASTIQUE ET ESTHETIQUE ANNALES DE CHIRURGIE PLASTIQUE ESTHETIQUE
ANNALES DES TELECOMMUNICATIONS-ANNALS OF TELECOMMUNICATIONS ANNALS OF TELECOMMUNICATIONS-ANNALES DES TELECOMMUNICATIONS
APPLIED AND COMPUTATIONAL MATHEMATICS COMPUTATIONAL APPLIED MATHEMATICS
ASTRONOMY AND ASTROPHYSICS ASTRONOMY ASTROPHYSICS

Les premiers journaux sont les suivants:
830 PAN AFRICAN MEDICAL JOURNAL
360 ACTA CRYSTALLOGRAPHICA SECTION E-STRUCTURE REPORTS ONLINE
358 PHYSICAL REVIEW
299 FEUILLETS DE RADIOLOGIE
289 EUROPEAN PHYSICAL JOURNAL
281 ANNALES D UROLOGIE
280 JOURNAL FRANCAIS D OPHTALMOLOGIE
268 ANNALES DE CHIMIE-SCIENCE DES MATERIAUX
256 ARCHIVES DE PEDIATRIE
255 DIABETES METABOLISM
248 PRESSE MEDICALE
245 PHYSICS LETTERS
223 JOURNAL OF AFRICAN EARTH SCIENCES
219 JOURNAL OF MAGNETISM AND MAGNETIC MATERIALS
208 ANNALES DE DERMATOLOGIE ET DE VENEREOLOGIE
202 ANNALES FRANCAISES D ANESTHESIE ET DE REANIMATION
191 COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE II

3.2 Harmonisation des Auteurs

Dans la majorité des bases bibliographiques, le champ Auteur est très souvent source d'erreurs (homonymes, fautes d'orthographe dans les noms, prénoms entiers ou simples initiales, inversions entre le nom et le ou les prénoms, inversions de lettres, doublement de lettres, pollutions de tous ordres : éditeurs, préfaceurs, directeurs, traducteurs, liens vers les adresses, …). Pour toutes ces raisons, il est nécessaire d'envisager un nettoyage puis un traitement morphologique poussé afin de

déterminer les correspondances les plus vraisemblables. Un dictionnaire de synonymes est issu de ce traitement, il doit être validé avant d'être utilisé. Voici un exemple des correspondances potentielles détectées par Tétralogie dans le corpus (Maroc/WoS).

AARAB, A.	AARAB, A
AARAB, L.	AARAB, L
AARAB, M.	AARAB, M
AASSIF, E.	AASSIF, E
AASSIF, EH	AASSIF, EL HOUCEIN
AASSIF, E. H.	AASSIF, EL HOUCEIN
AASSIF, ELHOUCEIN	AASSIF, EL HOUCEIN
AATIQ, A.	AATIQ, A
AAZZAB, B.	AAZZAB, B
ABABOU, A.	ABABOU, A
ABABOU, K.	ABABOU, K
ABABOU, M.	ABABOU, M
ABABOU, MOHAMED.	ABABOU, MOHAMED
ABABOU, R.	ABABOU, R
ABABSA, FAKHREDDINE	ABABSA, FAKHR-EDDINE
ABADABENDIB, M	ABADA-BENDIB, M
ABADA, REDALAH	ABADA, REDA LAH

3.3 Nombre de publications par Auteur

Les synonymies conservées sont alors prises en compte dans le calcul des fréquences de publication et dans toute matrice de croisement sur les auteurs. Ci-dessous, la liste des auteurs les plus prolifiques.

692	CHEN, S	592	EIGEN, G	585	BRANDT,A
640	YU, J	591	DICIACCIO, L	585	BESSON, N
631	MEYER, C	590	GIORGI, FM	585	BENJAMIN, DP
626	WANG,H	589	DAVIDEK, T	585	BARREIRO, F
622	LIU,M	588	DAM, M	584	EINSWEILER, K
622	BILOKON, H	588	CLEMENT, C	584	BURDIN, S
621	WANG, J	588	CLARK, A	584	BUANES, T
621	GABRIELLI, A	588	BOSMAN, M	584	BECK, HP
614	YANG,H	587	DE, K	584	ADYE, T
605	BENCHEKROUN, D	587	CATINACCIO, A	583	DI GIROLAMO, B
602	COLLOT, J	586	COSTANZO, D	583	DAI, T
598	DELMASTRO, M	586	CHEVALIER, L	583	CHAKRABORTY, D
596	DJAMA, F	586	CALVET, D	583	CAVALLISFORZA, M
596	CARLI, T	586	ARABIDZE, G	583	CASADEI, D
595	CHEN, H	585	DOBOS, D	583	BLANCHARD, JB
595	BOURDARIOS, C	585	DITTUS,F	583	BETHKE, S
595	ALEKSA, M	585	DERUE, F	583	BERINGER, J
594	BOONEKAMP, M	585	CORNELISSEN, T	583	BARLOW, N
593	CARMINATI, L	585	COOKE, M	583	BAKER, OK
593	ALONSO, A	585	CHOURIDOU, S	583	ALBRAND, S

3.4 Evolution du nombre de publications

		2010-11	2012-13	2014-15	2016-17
1	eigen,_	63	200	190	138
2	alonso,	62	196	194	138
3	benchek2	69	194	187	138
4	chen,_h	64	196	191	137
5	delmast	64	196	190	137
6	carli,_1	64	196	190	137
7	aleksa,	64	196	190	137
8	booneka	64	196	190	137
9	bourdar	64	195	190	137
10	davidek	63	196	190	137
11	dam,_m	63	196	190	137
12	catinac	63	196	190	137
13	bosman,	63	196	189	137
14	clark,_	62	196	190	137
15	arabidz	63	196	190	136
16	calvet,	63	195	190	137
17	chevali	62	196	190	137
18	chourid	62	196	190	137
19	dobos,_	62	196	190	137
20	djama,_	63	194	190	137
21	de,k	64	194	190	136
22	costanz	62	195	190	137
23	benjami	63	196	188	137
24	corneli	63	195	190	136
25	derue,_	64	195	190	135
26	dittus,	63	194	190	137
27	buanes,	59	199	189	137
28	einswei	62	196	190	136
29	besson,	60	196	190	137
30	cooke,_	60	196	190	137
31	beck,_h	63	196	189	135
32	beringe	59	196	190	138

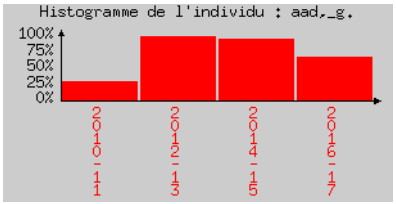
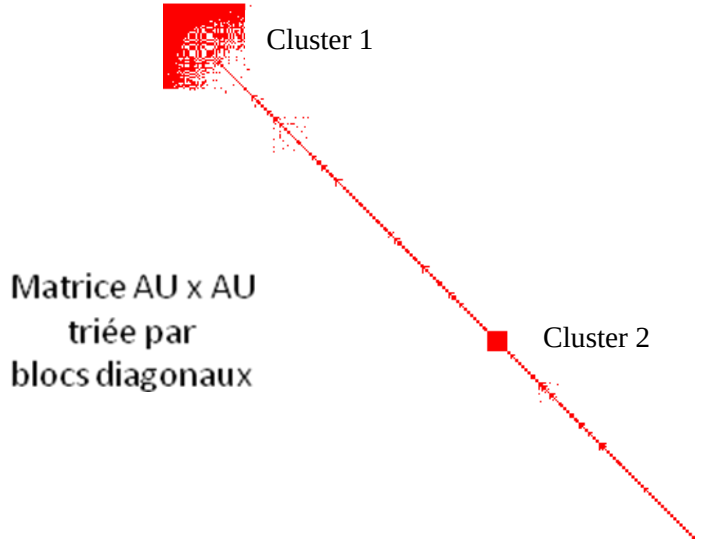


Figure 5 : évolution du nombre de publications entre 2010 et 2017

La première colonne représente le nombre de publications de l'auteur, la seconde est l'identifiant de l'auteur qui a été retenu après la phase de synonymie. Pour certains des auteurs, il y a un cumul des occurrences correspondant à plusieurs formes orthographiques rencontrées dans le corpus. N'est toujours pas réglé le problème des vrais homonymes (deux personnes différentes ayant exactement le même identifiant (nom, prénom ou initiale) dans la base Pascal. Heureusement ce type de collision peut être en partie corrigé en cours d'analyse, car l'auteur en question a souvent deux casquettes (deux domaines, deux équipes de collaborateurs, deux origines, deux groupes de journaux) qui n'ont que très peu ou pas de connexion par ailleurs. Il convient alors, soit d'éliminer cet auteur bicéphale pour une analyse macroscopique, soit le différencier dans le corpus en fonction de ses connexions avec son environnement. Ce travail fastidieux représente une des limites de notre approche et il ne pourra être évité que si les bases bibliographiques s'intéressent au problème (différentiation dès la saisie des articles), donc gestion d'une base de données des homonymes détectés.

3.5 Les équipes et de leurs relations

Pour cela, nous croisons les auteurs entre eux afin d'obtenir une matrice de cooccurrences qui sera filtrée, décomposée en classes connexes, elles mêmes triées par blocs diagonaux afin de faire apparaître la colonne vertébrale (équipes fortement structurées) de chaque classe. Le graphe global (plusieurs milliers d'auteurs) n'est pas manipulable, ni réellement utile, puisqu'il cumule des disciplines souvent très éloignées. Par contre, des clusters très marqués apparaissent sur le zoom de cette matrice correctement triée par blocs, en voici un exemple parmi d'autres.
Afin de montrer tout de même la structure de la recherche marocaine dans sa globalité, nous avons appliqué une simplification au graphe initial : nous n'avons gardé que les auteurs ayant 5 publications ou plus (soit une par an au moins) et nous avons négligé les liens à 1 (une seule publication en commun avec un autre auteur). Le graphe comporte alors 1000 sommets et il peut être dessiné sans difficulté. On y remarque des équipes bien structurées (celle se trouvant sous forme de blocs diagonaux dans le zoom de la matrice), par contre les liens sont au moins de 2 publications (sinon le graphe n'est plus lisible).



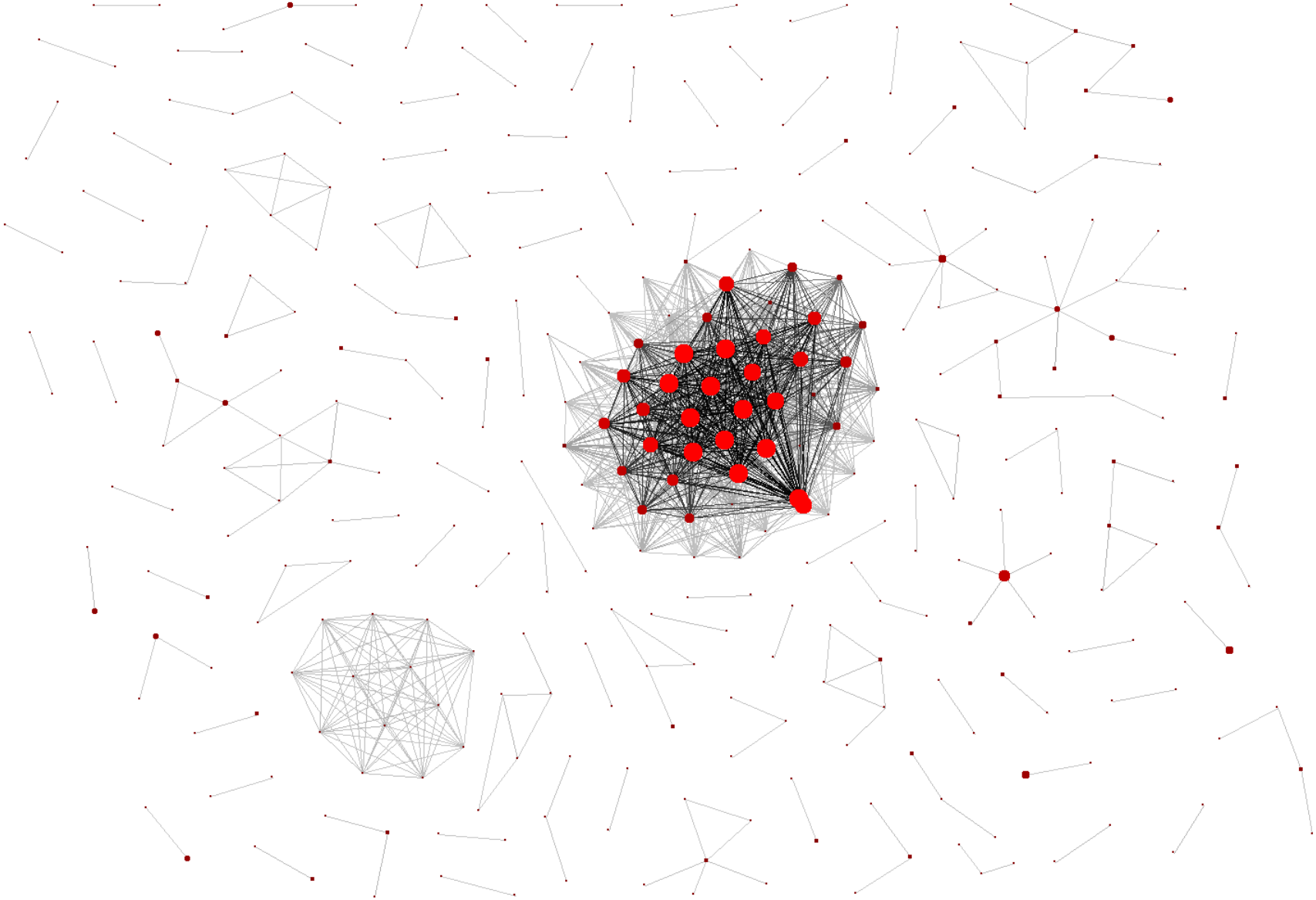


Figure 6 : Graphe et zoom de la matrice AU X AU pour les auteurs ayant plus de 10 publications.

4 ANALYSE DES COLLABORATIONS INTERNATIONALES

4.1 Difficultés de l'étude sur les pays

Les pays sont présents dans le champ Adresse (AD:) mais ils doivent être détectés parfois de façon indirecte
Par exemple :

- Guadeloupe et Martinique apparaissent sans la France
- Un état américain apparaît sans Etats-Unis
- Moroco pour Maroc et non pas Morroco, …
- On ne peut pas « couper » au point car il fait partie de certains Pays : U.S.A, U.K., …

Ces erreurs ou omissions sont corrigées par Tétralogie grâce, à nouveau, à des dictionnaires de synonymes :

AFGANISTAN	AFGHANISTAN	
AFRIQUE DU SUD	SOUTH-AFRICA	
ALABAMA.	USA	
ALABAMA	USA	
ALBANIA.	ALBANIA	
ALBANIE	ALBANIA	
ALBANIE.	ALBANIA	
ALEMANIA	GERMANY	
ALABAMA.	USA	
ALABAMA	USA	
ALBANIA.	ALBANIA	
ALBANIE	ALBANIA	
ALBANIE.	ALBANIA	
ALEMANIA	GERMANY	
ALGERIA.	ALGERIA	
ALGERIE	ALGERIA	
ALLEMAGNE	GERMANY	
AL USA.	USA	
AL USA	USA	
ANGLETERRE	UK	
ANTIGUA AND BARBUDA		
ANTIGUA-BARBUDA		
ARABIA	SAUDI-ARABIA	
ARABIA SAUDITA	SAUDI-ARABIA	
ARABIE-SAOUDITE	SAUDI-ARABIA	
ARG	ARGENTINA	
ARG.	ARGENTINA	

Le champ « Adresse » ainsi synonymé est ensuite filtré pour ne garder que des noms de pays valides. Pour cela nous utilisons un dictionnaire de pays préétabli qui permettra ensuite de dresser des cartes géostratégiques, nous en donnons ci-dessous le début :

AFGHANISTAN
ALBANIA
ALGERIA
ANGOLA
ANTIGUA-BARBUDA
ARGENTINA
ARMENIA
AUSTRALIA
AUSTRIA
AZERBAIJAN
BAHRAIN
BANGLADESH
BARBADOS
BELARUS
BELGIUM
BELIZE
BENIN
BHOUTAN
BOLIVIA
BOSNIA
BOTSWANA
BRAZIL
BRUNEI
BULGARIA
BURKINA-FASO
BURUNDI
CAMBODGE
CAMEROON
CANADA
CTRL-AFRICAN-REP
CHAD
CHILE
CHINA
COLOMBIA
CONGO
CONGO-PEOPL-REP

4.2 Relations internationales en nombre d'adresses par pays

Il peut y avoir jusqu'à 1 000 adresses pour un même document notamment en physique des particules. C'est pour cela qu'il y a, par exemple, plus d'adresses au Maroc que de documents. Les USA semblent donc être le premier collaborateur du Maroc avec 38 468 adresses, mais ils seront très loin en nombre de documents.

66455 MOROCCO
38468 USA
31189 FRANCE
22465 ITALY
14390 GERMANY
13471 JAPAN
12488 UK
11032 SPAIN
9441 CANADA
8069 RUSSIA
6054 CHINA
5778 TURKEY
5623 PORTUGAL
3938 SWITZERLAND
3800 SWEDEN
3030 BRAZIL
2893 ROMANIA
2854 CZECH-REPUBLIC
2747 NETHERLANDS
2747 AUSTRALIA
2715 SOUTH-AFRICA
2588 GREECE
2365 POLAND
2365 POLAND
2263 ISRAEL
2144 ARGENTINA
1834 TAIWAN
1537 NORWAY
1478 CHILE
1399 HONG-KONG
1380 SLOVAKIA
1358 SLOVENIA
1328 BELARUS
1176 BELGIUM
1151 SERBIA
1119 TUNISIA
1055 AUSTRIA
1001 INDIA
954 SAUDI-ARABIA
936 AZERBAIJAN
936 ALGERIA
922 DENMARK
849 HUNGARY
792 COLOMBIA
680 ARMENIA
563 MALAYSIA
435 EGYPT
426 IRAN
425 PAKISTAN
397 MEXICO
278 SOUTH-KOREA
257 UAE
244 FINLAND
230 SINGAPORE
220 BULGARIA
211 ETHIOPIA
198 JORDAN
189 SENEGAL
180 LEBANON
177 CAMEROON
151 NIGERIA
148 NEW-ZEALAND
112 ESTONIA
102 IRELAND
98 VENEZUELA
96 PHILIPPINES
95 VIETNAM
95 QATAR
92 KENYA
84 THAILAND
84 SYRIA
83 UKRAINE
81 CROATIA
80 KUWAIT
78 COTE-IVOIRE
73 PERU
73 BENIN
72 MALI
66 BURKINA-FASO
64 LUXEMBOURG
63 OMAN
60 GHANA
56 YEMEN
56 MAURITANIA
54 INDONESIA
52 URUGUAY
51 NIGER
49 SUDAN
48 TANZANIA
46 MADAGASCAR
45 BAHRAIN
42 UGANDA …

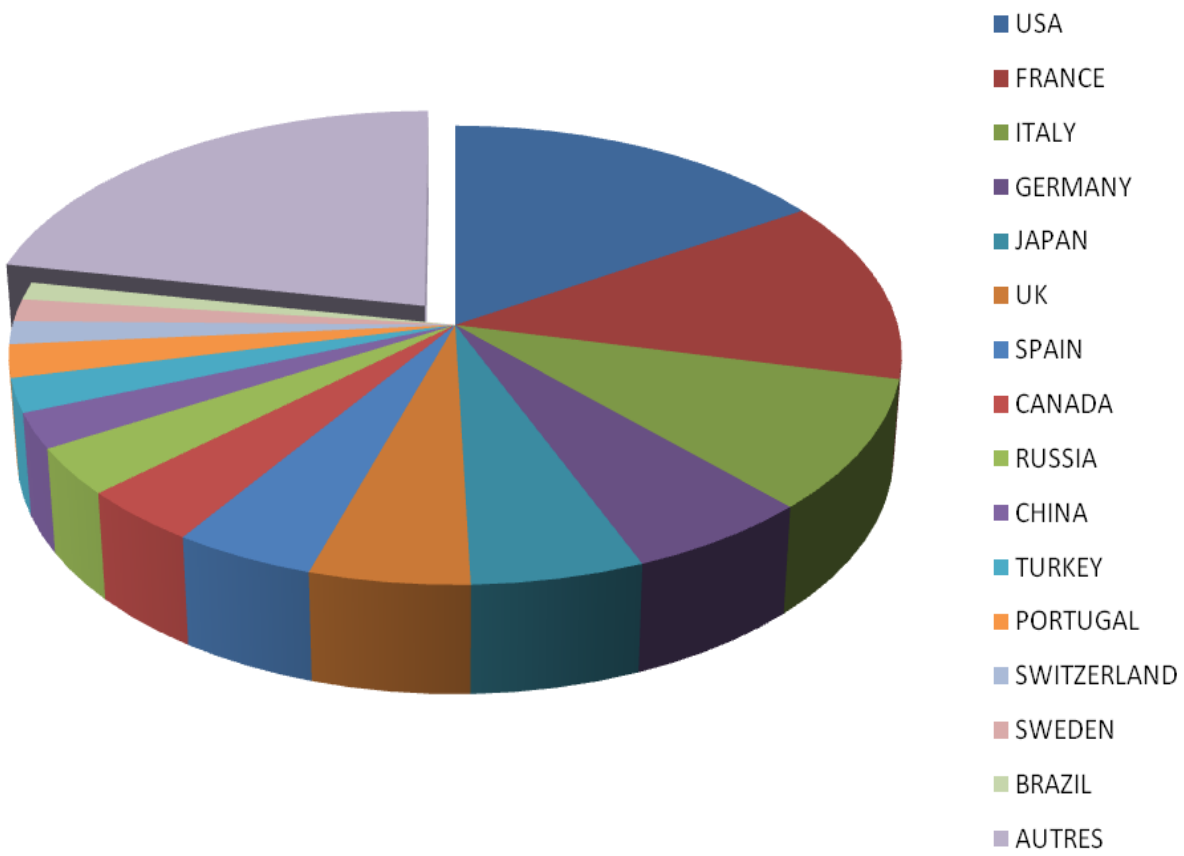


Figure 7 : répartition des collaborations en nombre d'adresses pour les principaux pays (sur représentation des USA).

4.3 Relations internationales en nombre de documents cosignées (période 2010-2017)

Si nous nous référons au nombre de publications cosignées par le Maroc avec chaque pays, la France passe en tête et les USA ne sont plus que 3^{ième}, l'Espagne venant s'intercaler en 2^{ième} position. Les pays européens sont bien représentés, hors Europe, la Chine, la Turquie et le Japon sont dans le Top 15. Il est possible de calculer le taux de progression de chaque pays, ainsi que le taux de progression du Maroc à l'international.

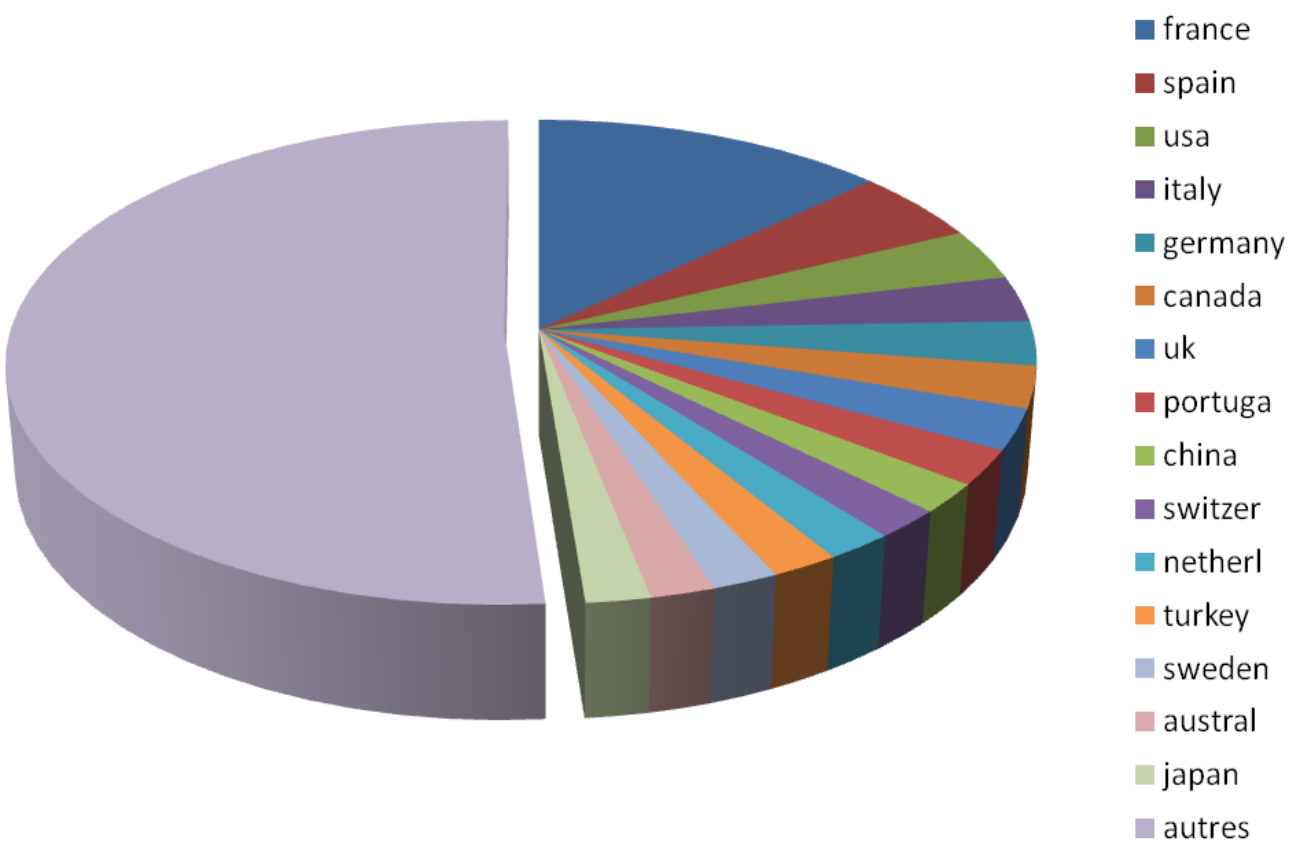


Figure 8 : répartition des collaborations internationales du Maroc pour les 8 dernières années (en nombre de documents cosignés)

4.4 Cartes cumulant les quatre périodes retenues (8 dernières années)

Dans cette carte : en grenat les pays n'ayant jamais collaboré avec le Maroc, en orange, ceux qui n'ont pas collaboré lors de la période 2010-2017 mais ayant eu des collaborations antérieures. Les différentes nuances de vert montrent l'importance de la collaboration sur la période, l'échelle n'est pas linéaire afin de faire ressortir les collaborations faibles comme pour l'Afrique, l'Amérique centrale, certains pays d'Asie ou d'Océanie.

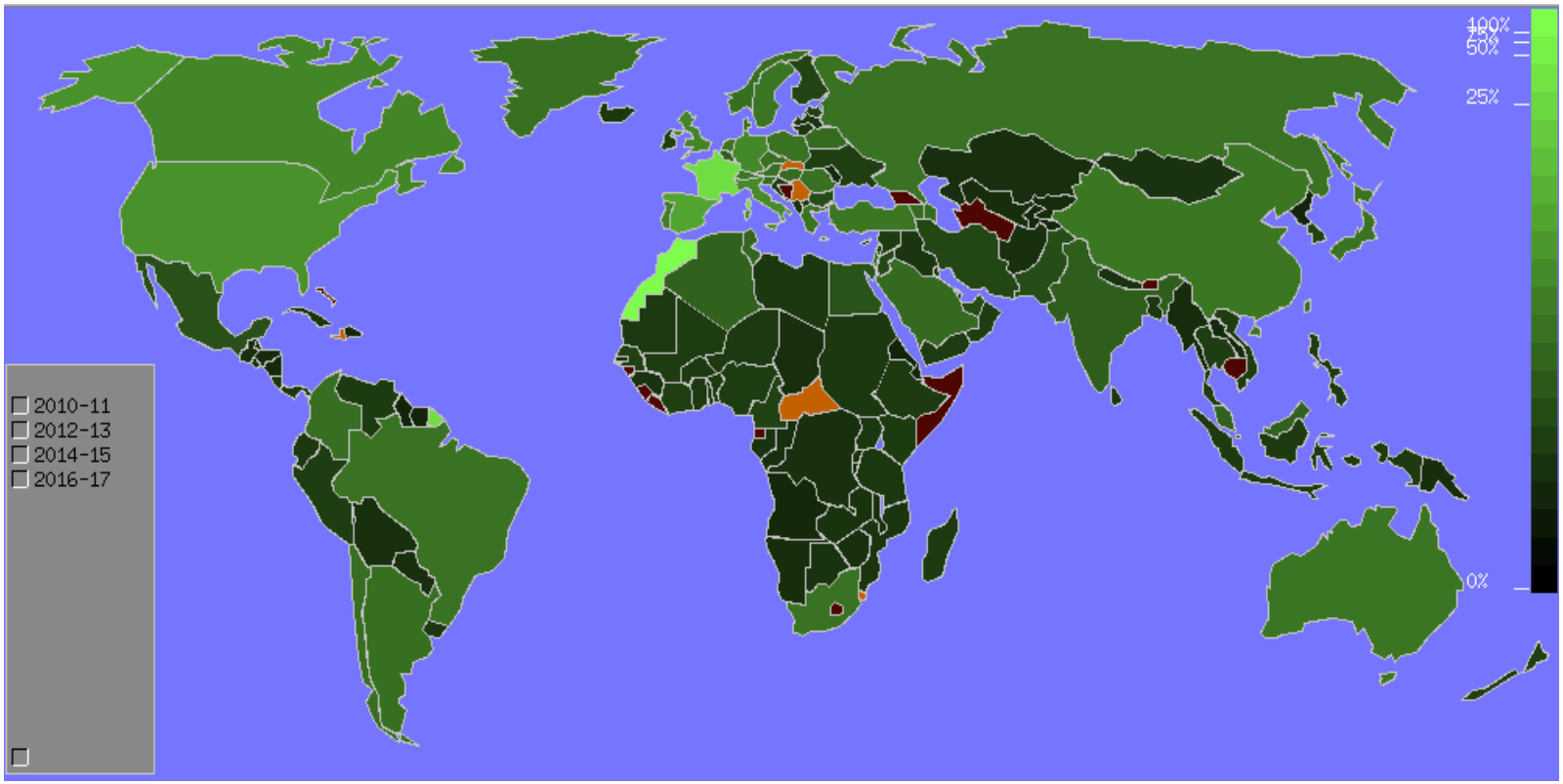


Figure 9 : cartes illustrant l'intensité des collaborations internationales du Maroc sur la période 2010-2017.

4.5 Evolution entre les cartes des quatre périodes

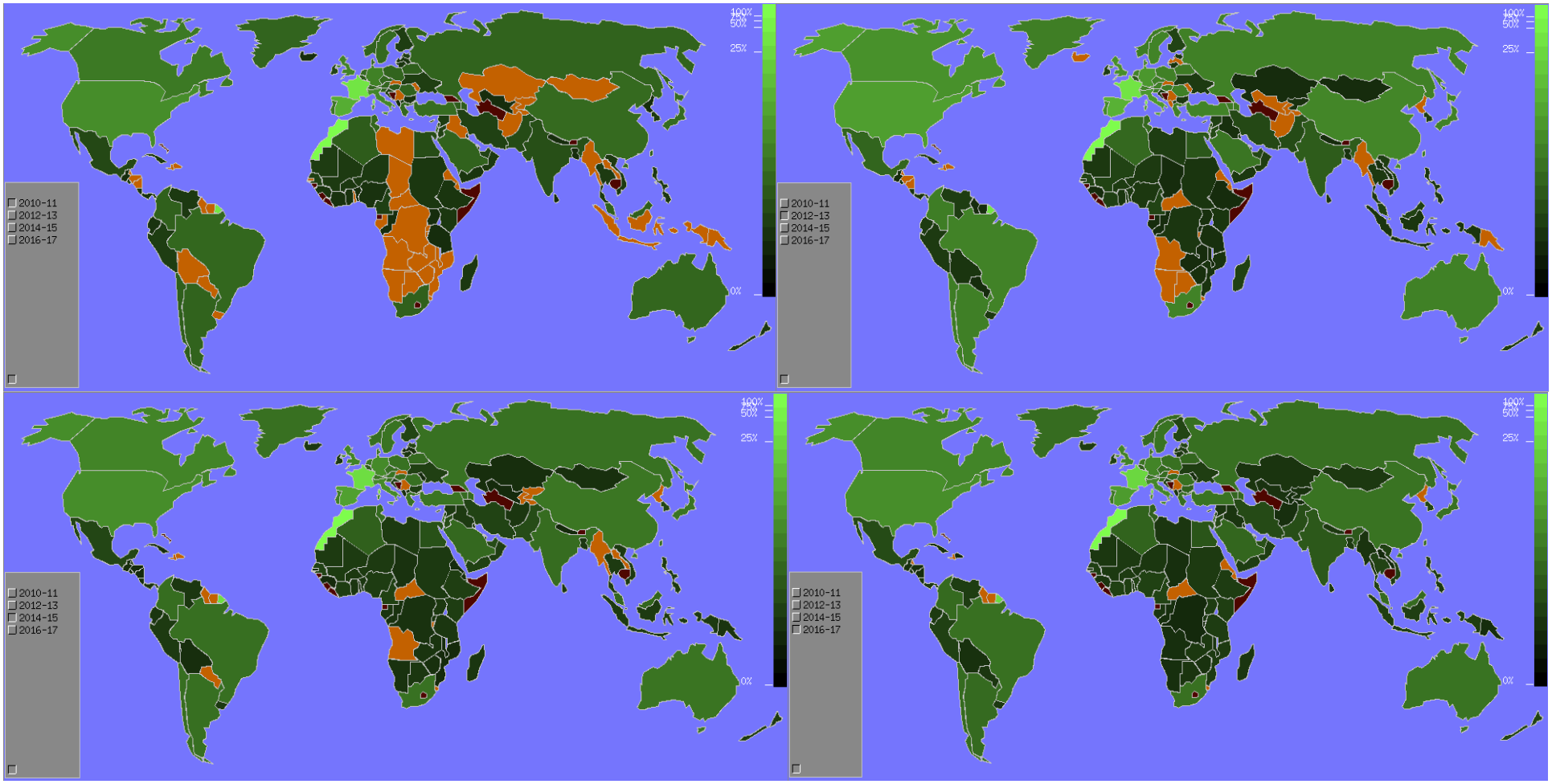


Figure 10 : cartes illustrant l'évolution des collaborations internationales du Maroc sur les 8 dernières années.

Nous utilisons, ici, les fonctionnalités du logiciel graphique GéoECD [S. KAROUACH 2003] qui permet de générer automatiquement des cartes géographiques interactives à partir des tableaux de données produits par « Tétralogie ». Ces cartes sont entièrement manipulables par tout utilisateur local ou distant et permettent donc, dans des temps très courts, de communiquer des informations géostratégiques via le réseau avec la possibilité de les retravailler à distance. Nous avons ici décomposé le temps en 4 périodes de 2 ans (2010-11, 2012-13, 2014-15, 2016-17). Nous présentons, sous forme de 4 cartes, les 4 périodes choisies. La sélection des colonnes du tableau de données s'effectue à gauche en cliquant sur le bouton correspondant à chaque période. Il est aussi possible de cumuler plusieurs colonnes et donc plusieurs périodes.

Pour un tableau ayant un grand nombre de colonnes, les choix de cartes sont quasiment infini : choix des colonnes, du codage des couleurs, des pondérations (par la population, le pnb, la surface utile, …), des zooms (continents, zones géographiques, G7, choix de l'utilisateur, …).
Il ressort de cette analyse que les collaborations internationales récentes sont en nette progression, notamment avec le reste de l'Afrique, l'Europe centrale, l'Amérique centrale, l'Amérique du sud ainsi que l'Indonésie. Pour les autres pays, il y a une certaine stabilité, la France restant toujours en tête.

4.6 Répartition des recherches par région marocaine

Pour pouvoir déterminer la ou les régions du Maroc d'où est issue une recherche, nous avons établi une correspondance entre les principales villes du Maroc et leur propre région. Afin de n'oublier personne, après avoir appliqué la correspondance, nous avons vérifié s'il ne restait pas de ville sans affectation.

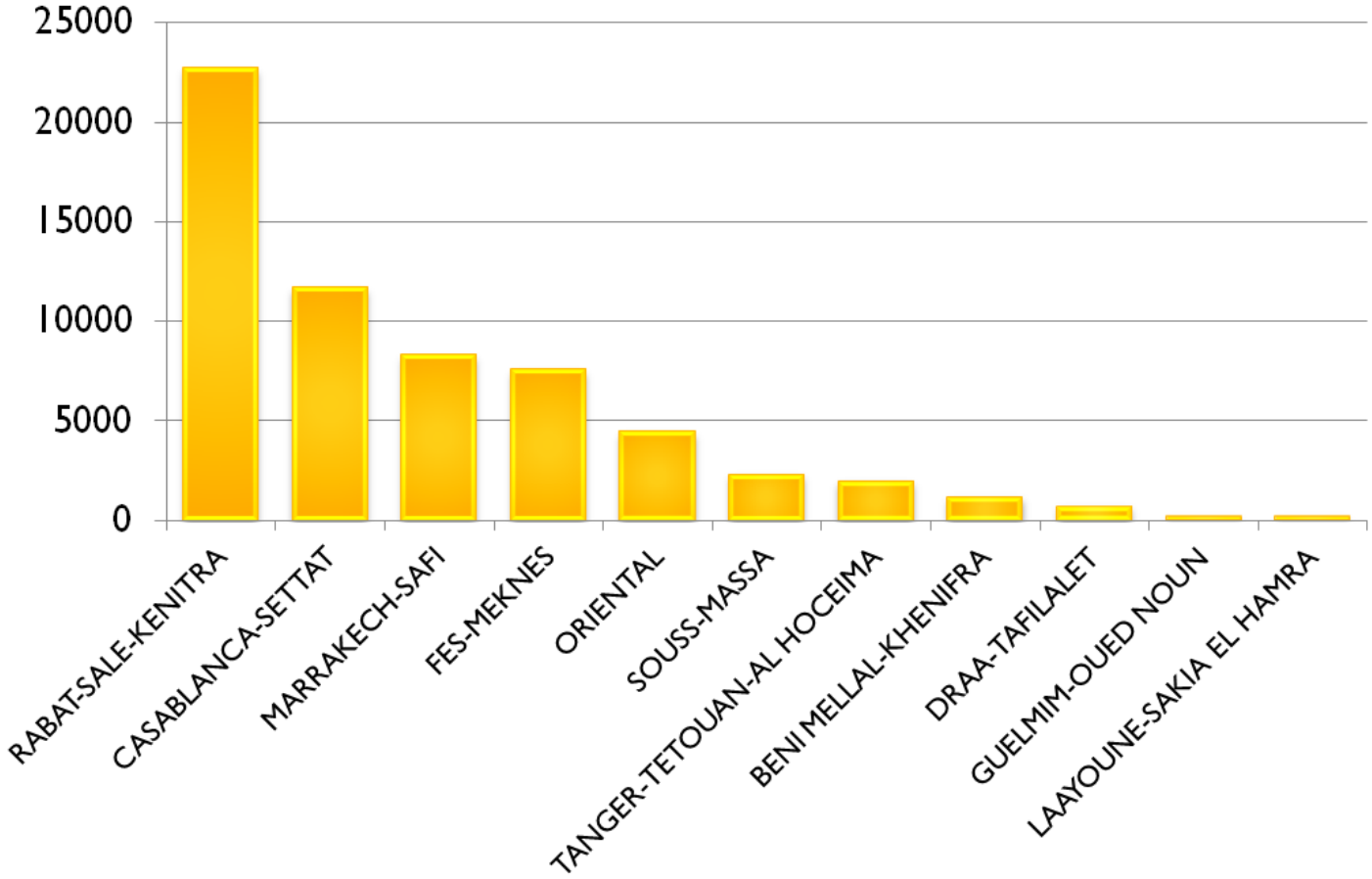


Figure 11 : répartition des recherches marocaines en fonction des régions.

5 LES SIGNAUX FORTS DE LA RECHERCHE MAROCAINE

5.1 Exploitation du champ mots-clés

Pour les champs mots-clés qui ont été fusionnés, il est nécessaire de générer des correspondances morphologiques dues aux pluriels, tirets, inversions, coupures variations orthographiques, mots outils manquants, … Pour les mots-clés ce dictionnaire de synonymes fait près de 800 lignes. En voici un extrait :

1M HCL 1 M HCL
3-DIPOLAR CYCLOADDITION 3-DIPOLAR CYCLOADDITIONS
3 TRIAZOLE 3-TRIAZOLE
AB INITIO AB-INITIO
AB INITIO CALCULATION AB-INITIO CALCULATION
AB INITIO CALCULATIONS AB-INITIO CALCULATION
AB-INITIO CALCULATIONS AB-INITIO CALCULATION
AB INITIO STUDY AB-INITIO STUDY
ACCELERATION OF CONVERGENCE CONVERGENCE ACCELERATION

Le champ mots-clés est divisé en 2 : les mots-clés de la base et les mots-clés des auteurs. Ces deux indexations peuvent être étudiées séparément ou bien cumulées, dans ce cas on ne déclare pas, dans les métadonnées, la seconde balise. Voici les termes, après synonymies, les plus fréquemment rencontrés dans le corpus :

1758 MOROCCO
322 ADSORPTION
216 CORROSION
200 X-RAY DIFFRACTION
185 CRYSTAL STRUCTURE
183 COMPONENT
181 TUBERCULOSIS
181 SURGERY
162 CLOUD COMPUTING
158 TREATMENT
152 EPIDEMIOLOGY
147 CHILD
143 INHIBITION
142 SECURITY
136 OPTIMIZATION
135 CORROSION INHIBITION
129 DIAGNOSIS141 XRD-
126 NORTH AFRICA
125 STABILITY
122 CLASSIFICATION
120 MODELING
119 DFT
118 SIMULATION
114 MRI
113 THIN FILMS
107 MAGNETIC PROPERTIES
105 STEEL
105 PHASE DIAGRAM
104 MILD STEEL
102 ESSENTIAL OIL
102 CHILDREN
99 HADRON-HADRON SCATTERING
99 GENETIC ALGORITHM
98 MONTE CARLO SIMULATION
95 HEAVY METALS
93 OPTICAL PROPERTIES
91 WIRELESS SENSOR NETWORK
89 RAMAN SPECTROSCOPY
89 COPPER
88 AFRICA
84 KIDNEY
83 PHASE TRANSITION
83 HEAT TRANSFER
82 PREVALENCE
81 EIS
80 PROGNOSIS
79 CLUSTERING, …

5.2 EXTRACTION DES SIGNAUX FAIBLES

Afin de faire apparaître les signaux faibles, nous réalisons, dans un premier temps, une analyse factorielle des correspondances (AFC) sur la matrice qui croise les mots-clés cumulés et le temps exprimé en 4 périodes : MC x PP. Une visualisation de la carte factorielle des périodes (PP) permet d'isoler 2016-17 dans un coin de la carte (ici en haut à droite), une exportation de l'azimut ainsi déterminé vers la carte des mots-clés (MC) permet de détecter tous ceux qui sont typiques de la période la plus récente 2016-17 et ce, quelque soit leur nombre. Une capture à la souris permet ensuite d'en extraire la liste qui est récupérée dans un filtre. Dans un second temps, il est possible de croiser ces termes émergents entre eux afin de regarder, comme précédemment, s'ils s'organisent en clusters significatifs. Si c'est le cas, on a détecté les signaux faibles du moment. Chaque cluster est ensuite listé et croisé avec les autres variables de la base (auteurs, laboratoires, pays, mots-clés non émergents, journaux, …) afin d'en valider la pertinence. En effet, si un laboratoire prestigieux ou une équipe de renom est lié à un signal faible, celui-ci est digne d'intérêt. Idem s'il s'agit de journaux de premier plan ou de pays donc la recherche est réputée. Bien souvent les experts sont déstabilisés par ce type d'information, car ils ne sont, bien entendu, ni à l'origine de cette recherche, ni dans le petit cercle d'initiés qui, éventuellement, est au courant de son existence. La seule façon de les convaincre est de leur en montrer l'origine. La matrice triée ci-dessous permet de visualiser le long de la diagonale les principaux clusters constitués de termes émergents, en haut à gauche sa partie connexe, en bas à droite les clusters isolés.

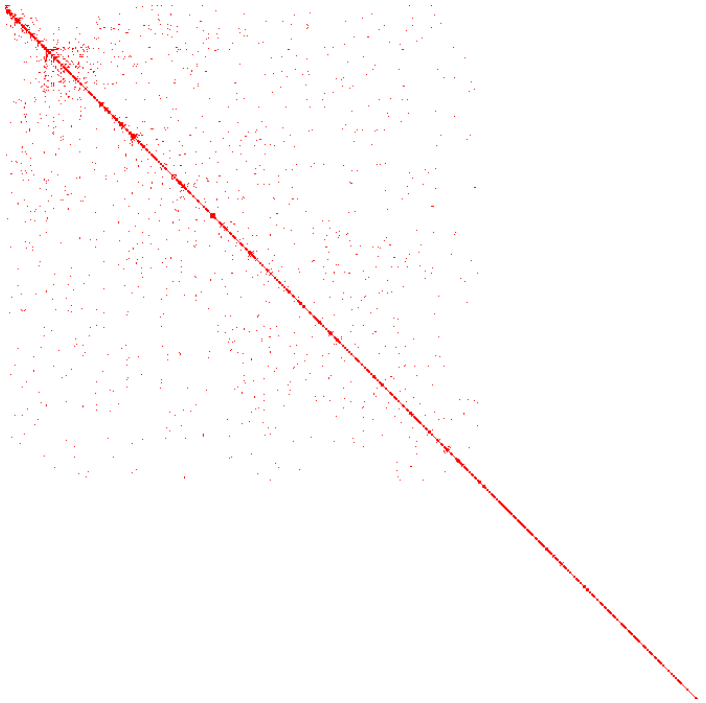
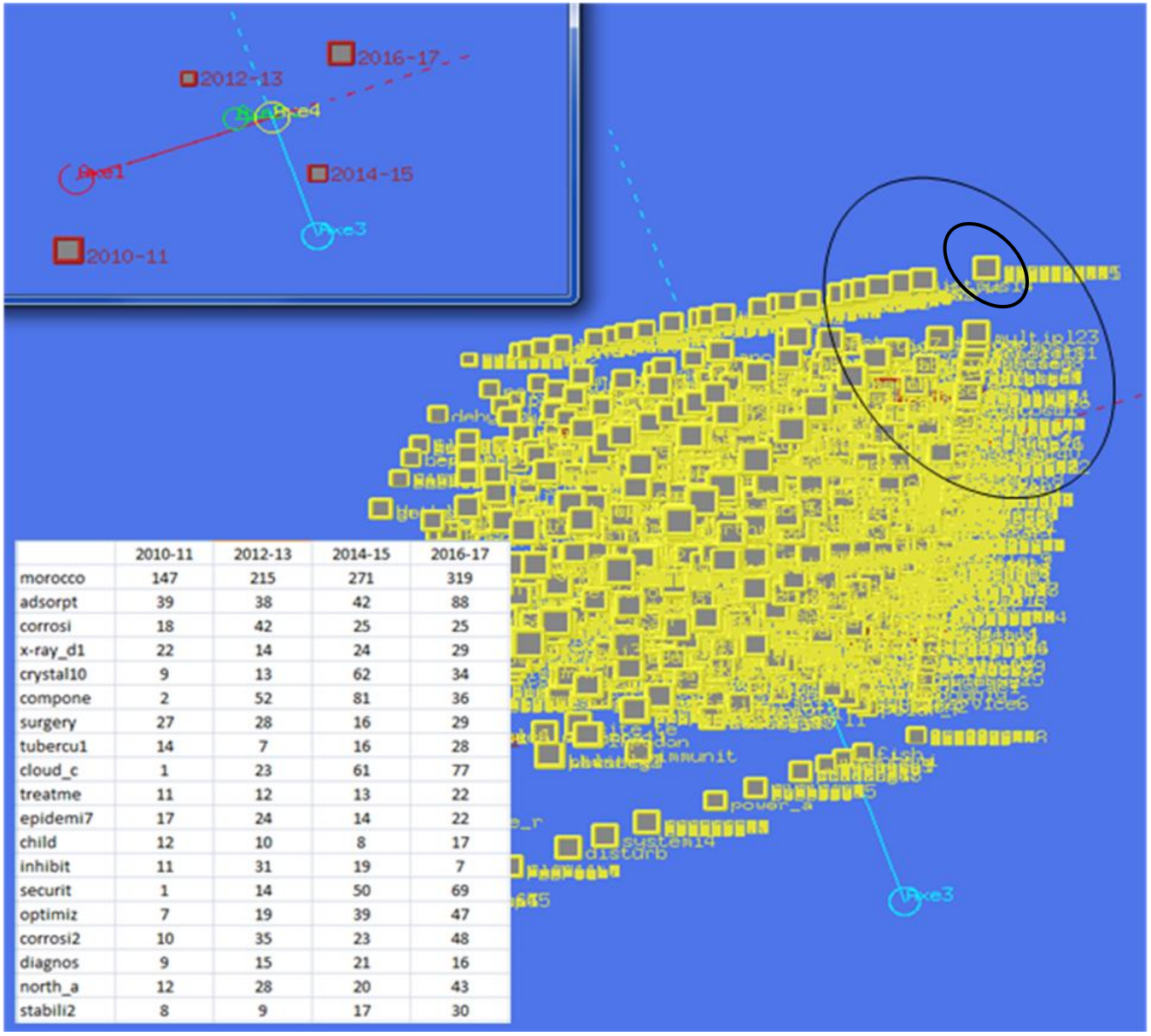


Figure 12 : tri par blocs diagonaux de la matrice croisant les termes émergents, chaque bloc est un signal faible.



	2010-11	2012-13	2014-15	2016-17
morocco	147	215	271	319
adsorpt	39	38	42	88
corrosi	18	42	25	25
x-ray_d1	22	14	24	29
crystal10	9	13	62	34
compone	2	52	81	36
surgery	27	28	16	29
tubercu1	14	7	16	28
cloud_c	1	23	61	77
treatme	11	12	13	22
epidemi7	17	24	14	22
child	12	10	8	17
inhibit	11	31	19	7
securit	1	14	50	69
optimiz	7	19	39	47
corrosi2	10	35	23	48
diagnos	9	15	21	16
north_a	12	28	20	43
stabili2	8	9	17	30

Figure 13 : AFC sur l'évolution des mots-clés et présence par période de 2 ans (en haut à droite les signaux faibles)

5.1 EXEMPLES DE SIGNAUX FAIBLES

Voici un extrait des signaux faibles obtenus en croisant les descripteurs en émergence (période 2016-17). Chaque cluster peut être extrait directement sur la matrice triée, ou depuis le graphe en utilisant la fonction de transitivité depuis l'élément central du cluster.

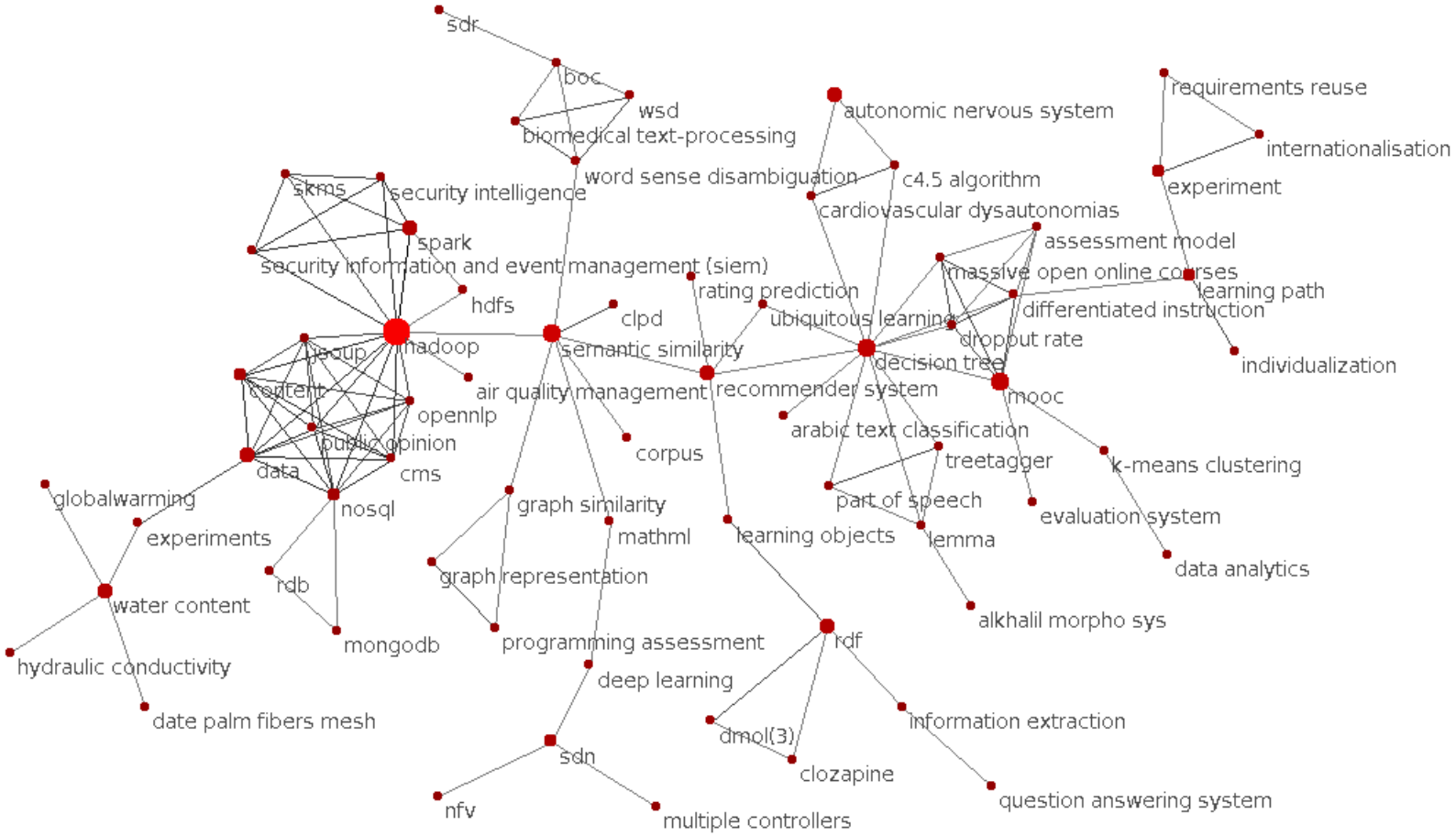


Figure 14 : réseau sémantique du signal faible autour de «hadoop» (filtrage à 75% sur 2016-17).

Ci-dessus, seuls les termes présents à plus de 75% sur la dernière période (2016-17) sont conservés. Si nous descendons le seuil à 66%, des termes moins spécifiques à cette période sont pris en compte comme Big Data, qui est tout de même en très forte croissance. Par contre la transitivité dans le graphe est moins poussée afin de garder une certaine lisibilité.

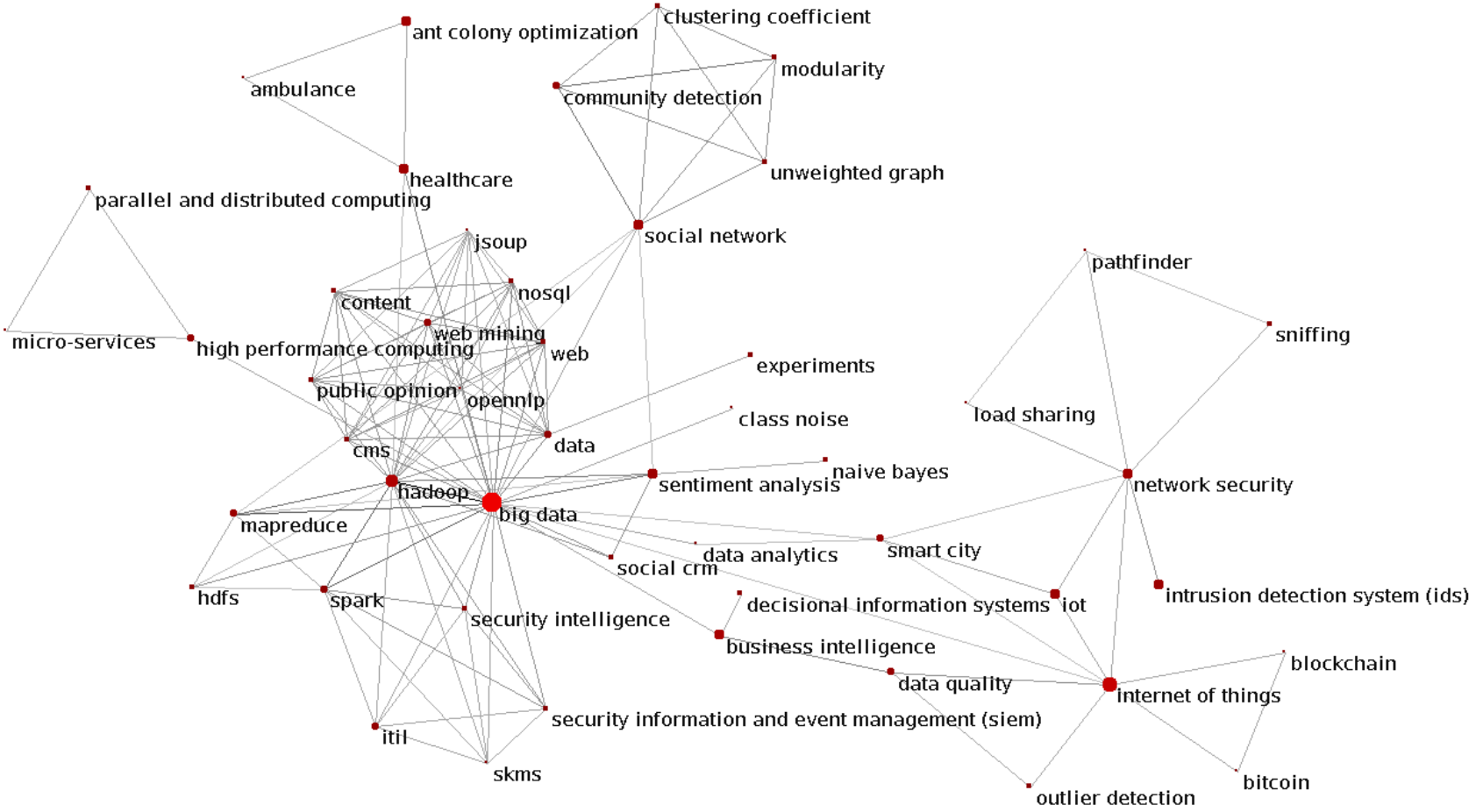


Figure 15 : réseau sémantique du signal faible «Big Data» (Filtrage à 66% sur 2016-17).

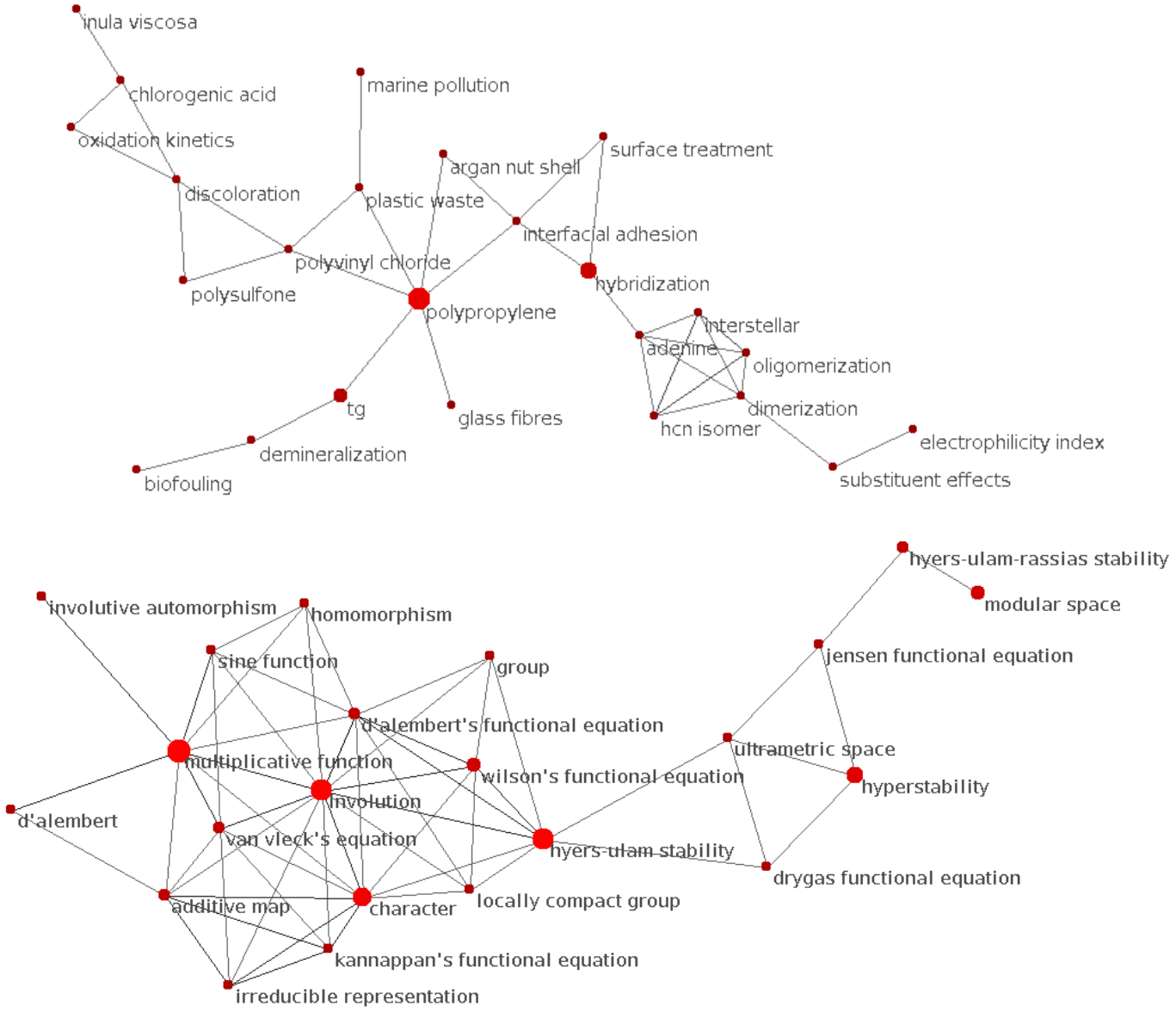


Figure 16 : réseaux sémantiques des signaux faibles «polypropylene» et «involution».

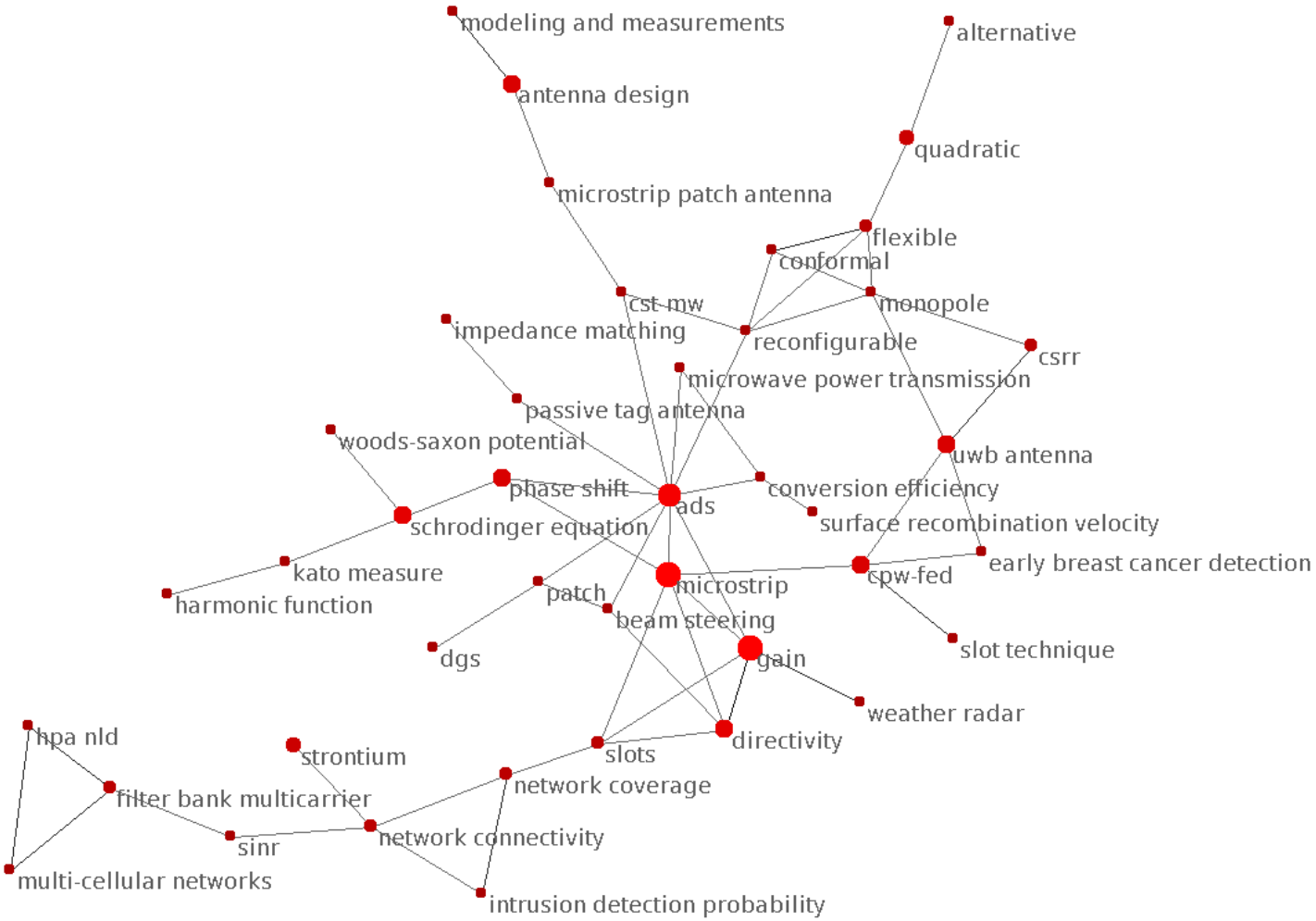


Figure 17 : réseau sémantique du signal faible «Microstrip» (Filtrage à 66%).

CONCLUSION

- Cette étude est limitée:
  - Une seule base: le WoS de Thomson
  - L'évolution n'est étudiée que sur 2010-2017
- Elle est assez rapide à mener
  - L'investissement initial est réutilisable
  - La mise à jour en est facilitée
- Elle est consultable à distance
  - Connexion ssh à Tétralogie
  - Compilation pour le portail Xplor (BD Mysql)
- Des focus sont recommandés sur chaque discipline
- Retombées possibles:
  - Management de la recherche
  - Evaluation des laboratoires et des filières
  - Rapprochement des équipes
  - Mise en commun de moyens et de réseaux de contacts, …

Nous avons volontairement limité la présentation de cette étude pour pouvoir présenter les principales méthodes utilisées en ce concentrant sur des résultats synthétiques obtenus dans un temps très limité. Le corpus constitué à partir du WoS peut encore livrer de nombreuses informations utiles, soit globales, soit plus ciblées comme par exemple sur un laboratoire, une équipe ou un domaine de recherche bien précis. Comme ces données sont en ligne sur le serveur tetralogie.irit.fr, il est possible d'affiner à distance (via une connexion ssh) l'ensemble des analyses déjà produites et d'en réaliser d'autres notamment à partir des champs encore peu utilisés. Ce type d'analyse nous est régulièrement demandé par des grands laboratoires de recherche afin de caller au mieux leur politique à long terme : collaborations internationales, suivi des sujets porteurs, détection de nouveaux axes (signaux faible), évaluation de la recherche, gestion des abonnement aux revues, recherche de partenaires, mise à jour de l'indexation, facteur d'impact, cartographie des connaissances, aide à la mise au point d'ontologies du domaine, recherche d'information à partir des cartes sémantiques. Les applications sont nombreuses, mais ce qui compte c'est que sous un même formalisme et avec des traitements bien maîtrisés, il est possible d'envisager l'étude de pratiquement toutes les sources d'information électronique et même de les combiner pour s'approcher de l'exhaustivité et éviter ainsi certains biais des études mono source.

6 BIBLIOGRAPHIE

[1] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret, *"Information mining: use of the document dimensions to analyse interactively a document set. "* 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

[2] J.-L. Multon, G. Lacombe, B. Dousset, *"Analyse bibliométrique des collaborations internationales de l'INRA".* Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

[3] B. Dousset, S. Karouach, *"Collaboration interactive entre classifications et cartes thématiques ou géographiques. "* 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.

[4] J.-L. Multon, G. Lacombe, B. Dousset, *"Analyse bibliométrique des collaborations internationales de l'INRA"*. 9ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, 14-18 octobre 2002.

[5] J. Mothe, C. Chrisment, B. Dousset, S. Karouach, *"Représentation des documents textuels : étude d'un domaine à travers des publications"*. 5ième Congrès de la société française de recherche opérationnelle et d'aide à la décision ROADEF. (Avignon France), pp 130-131, 26-28 février 2003.

[6] S. Karouach, B. Dousset, *"Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grande taille"*. Workshop sur la recherche d'information : un nouveau passage à l'échelle, associé à INFORSID'2003, (Nancy France), 3-6 juin 2003.

[7] J. Mothe, C. Chrisment, B. Dousset, J. Alaux, *"DocCube : multi-dimensional visualization and exploration of large document sets*". Journal of the American Society for Information Science and Technologiy JASIST, Special topic section: "Web Retrieval and Mining". Guest Editor: Hsinchun Chen, 54(7), pp 650-659, March 2003.

[8] S. Karouach, B. Dousset, *"Analyse d'information relationnelle par des graphes interactifs de grandes tailles*". 4ièmes journées d'EGC (Extration et Gestion de Connaissances), Clermont Ferrand, 20-23 janvier 2004.

[9] C. Chrisment, B. Dousset, S. Karouach, J. Mothe, *"Information mining : extracting, exploring and visualising geo-referenced information*". Workshop on Geographic Information Retrieval, SIGIR 2004.