Session 1
Introduction a CorText Manager

CorText Training Sessions, Janvier 2021

Lionel VILLARD

Marc BARBIER

LISIS, IFRIS, INRAE, CorText, ESIEE Paris

Présentation de CorText

Constituer une **plateforme scientifique et technique** avec ses exigences propres de positionnement liées à son existence et à son devenir pour soutenir un **espace de recherche sur les infrastructures, les traces et les usages numérique de la sciences et de l'innovation en société**.

2008	Equipes INRA Praxis et TSV		GIS IFRIS Projet PHARE CNRS
2010	UR INRA SenS	Projet SAD ESR sur Plateforme	GIS IFRIS LABEX SITES
2015	UMR LISIS	Ingénieur INRA	LABEX SITES
2019	UMR LISIS	Ingénieur INRA Ingénieur ESIEE	RISIS-1 RISIS-2

Présentation de CorText

Marc Barbier
Member of CorText platform, Researcher at LISIS, Head of IFRIS

Antoine Schoen
Member of CorText platform, Researcher at LISIS, Senior lecturer at ESIEE Paris

Lionel Villard
Head of CorText platform, Researcher at LISIS, lecturer at ESIEE Paris

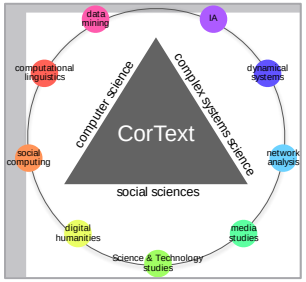
Patricia Laurens
Member of CorText platform, Researcher at CNRS and LISIS

Philippe Breucker
IT engineer from INRAE, LISIS. Technical Director of the Cortext Digital Platform. Web Designer and developer.

Bilel Benbouzid
Researcher, Senior lecturer at LISIS

Alexandre Hannud Abdo
Post-doctorant, LISIS

Juan-Pablo Ospina Delgado
IT Engineer, Cortext



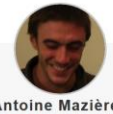
Luis-Daniel Medina
IT Engineer, Cortext

Pierre-Yves Bulot
IT Engineer Assistant, Cortext

Diego-Fernando Gómez Peña
IT Engineer, Cortext

Tatiana Andrea Sánchez Castaño
IT Engineer, Cortext

Joenio Marques da Costa
Research Software Engineer, Cortext



Antoine Mazières
Research scientist in the Computation Social Science team at Centre Marc Bloch

Constance De Quatrebarbes
Fondateur Présidente - DRISS (Digital Research In Science & Society)

Chloé Duloquin
Web Designer, Graphiste, Intégratrice web

Jean-Philippe Cointet
Associate Professor, Sciences Po Paris, Medialab

Guillaume Orsal
Computer engineer, data mining, web development and SEO

Cristian Martinez
PhD Engineer in Computer Science, NLP/Data Senior Consultant at Cogniteva

Nicolas Turenne
Assistant professor in data science, Beijing Normal University & Hong Kong Baptist University United International College

Tam Kien Duong
Data & design, Etalab

Nicolas Baya-Laffite
STSLab, Université de Lausanne

Loic Boudoulec
IT Engineer

Bertha Brenes
IT Engineer, Trainee, Cortext

Anis Arabi
Big data engineer

Nicolas Ricci
Web developper and data

Audrey Baneyx
Project manager Data science, Sciences Po Paris - médialab

Andrei Mogoutov
Bullescence

Élise Tancoigne
Researcher, University of Geneva, Switzerland

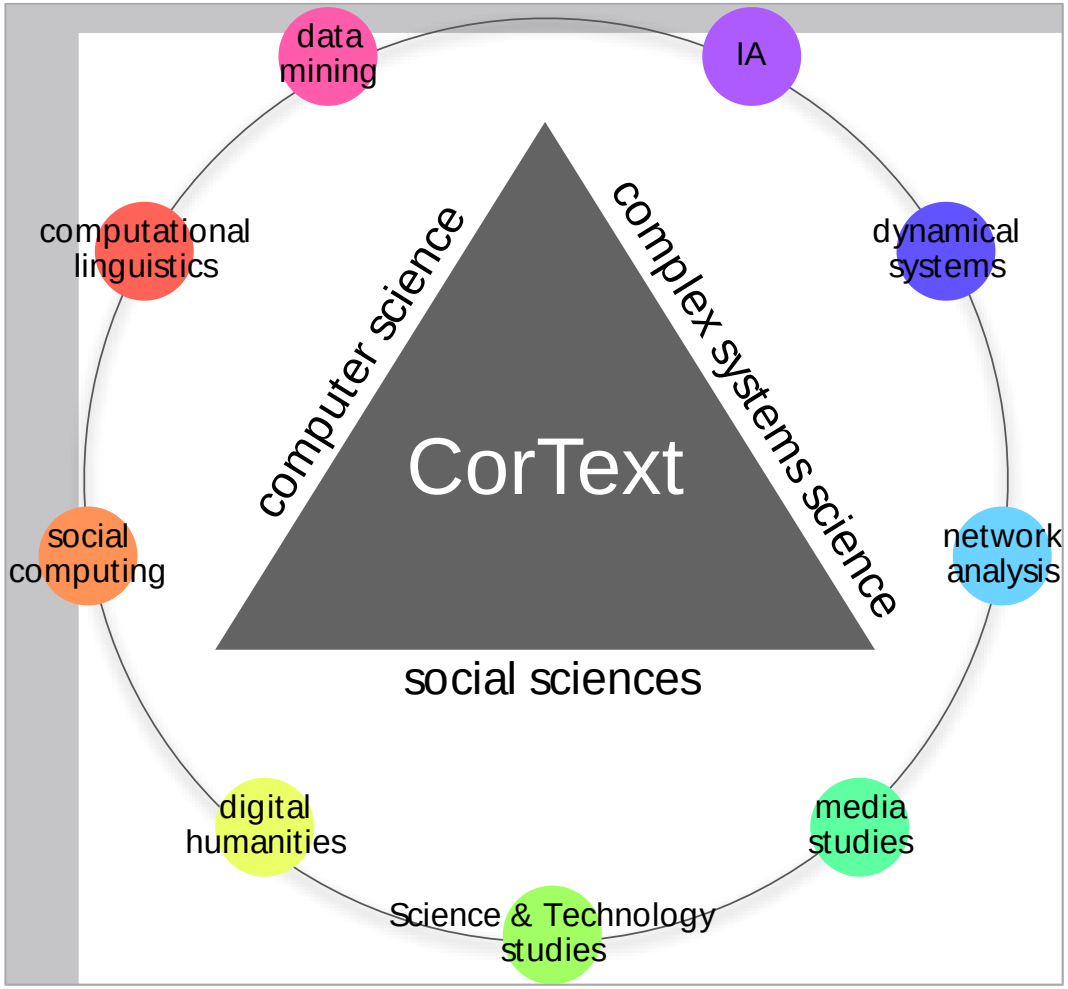
The origin of our world / what could be original?

Manuscrits > Imprimerie > Informatique > Hyperliens et traces numériques

Déluge de données

The origin of our world / what could be original?

* 1955, Eugene Garfield crée un répertoire interdisciplinaire pour les bibliothécaires regroupant les articles des principaux périodiques et leurs références (Garfield, 1955).
* 1958, fondation de l'Institute for Scientific Information (ISI) et première version papier du Science Citation Index (SCI) en 1963.
* 1963, Solla Price publie "Little science, big science" et impose l'idée d'une mesurabilité de la production scientifique reflet de loi sociales.



Measurement of text

Ressources

. Compétences INRA, ESIEE

. Financements: GIS IFRIS, LABEX SITES, MP SMaCH, PHARE CNRS

Projets de recherche EU RISIS, ANR PanBioptique, EMBRIC,

Partenariats: ISCPIF, MediaLab SciencesPo

Les axes de travail fréquent

- **Emergence** de domaine de recherche et d'innovation (bioénergie, nanotechnologie, biodiversité)
- **Controverses** et « hot topics » (pesticides, biologie de synthèse, sécurité alimentaire mondiale)
- **Cartographie socio-sémantique et relationnelle** des production de la recherche (publications, brevets, projets)
- Analyse du **web et des média-sociaux** pour caractériser les phénomènes de la Science et de l'Innovation en Société

La scientométrie

Aux origines de l'analyse de traces numériques : l'article scientifique

Les articles scientifiques comme source d'information ?

A ce titre, un **article scientifique** est considéré comme un indicateur important de la production de la recherche scientifique (mais pas le seul).

Les « **connaissances certifiées** » sont des connaissances qui ont été soumises à la critique des collègues et qui ont résistées à leurs objections (Callon, 1993).

Dés 1962, Derek de Solla Price identifie des lois générales caractérisant l'activité des scientifiques en appliquant aux articles scientifiques des **analyses quantitatives** (documents pour comprendre des dynamiques scientifiques et sociales).

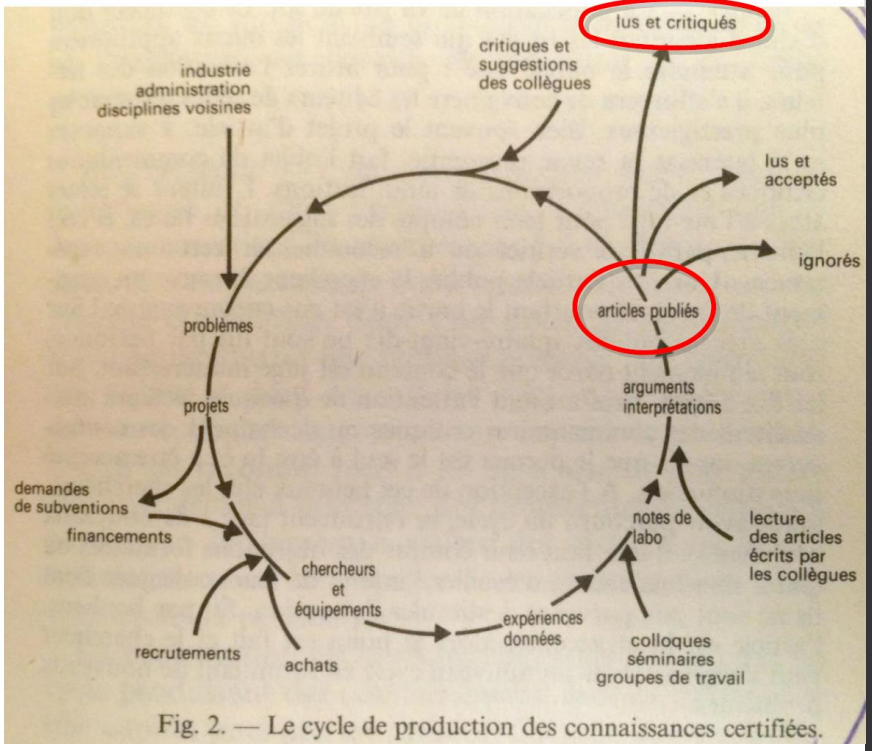


Fig. 2. — Le cycle de production des connaissances certifiées.

Les articles scientifiques comme source d'information ?

ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Research Policy 36 (2007) 893–903

research policy
www.elsevier.com/locate/respol

Journal

Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking☆

Titre : haut niveau de synthèse sur le contenu de l'article

Andrei Mogoutov [a,*], Bernard Kahane [b,c,1]

Auteurs : collaboration scientifique

[a] *AGUIDEL, 68 Bld de Port Royal, 75005 Paris, France*
[b] *LATTS (Laboratoire Territoires, Techniques et Sociétés), CNRS/UMLV/ENPC, École Nationale des Ponts et Chaussées, 6-8 avenue Blaise Pascal, Cité Descartes, Champs sur Marne, 77455 Marne La Vallée Cedex 2, France*
[c] *ISTM (Institut Supérieur de Technologie et Management), Cité Descartes, 93162 Noisy le Grand Cedex, France*

Adresses : institutions et géographie des auteurs

Available online 23 April 2007

Date de publication : dimension temporelle

Abstract

Nanotechnology, like other emerging technologies that increasingly characterize the dynamic of our era, makes specific demands on datamining to track and interpret efficiently what is happening, through publications and other scientific output. We here propose and describe a strategy based on an automated lexical modular methodology to overcome rapidly evolving content and classification problems, which may otherwise accommodate poor quality of data and expert bias, with potential dire consequences for interpretation, decision and strategy. The proposed methodology is based on an initial nanostring enriched and screened by eight subfields, automatically identified and defined through the journal inter-citation network density displayed in the initial core nanodataset. Relevant keywords linked to each subfield are then tested for their specificity and relevance before being sequentially incorporated to build a modular query. We then, as a first test, compare the database constructed using this methodology for years 2003 and 2005 with those obtained by other approaches previously used to cover and explore the nanotechnology dynamic. Finally, using the inherent transparency, portablity and replicability of our methodology, we offer, in order to help our initial query evolve and develop, a set of evaluation processes for tests by researchers in the nano field, other scientometric teams and intelligence experts involved in decision-making processes.

Résumé : contenu de l'article

Keywords: Datamining; Nanotechnology; Emergent technologies

Mots clefs des auteurs (vision synthétique de l'article par l'auteur) : notions, concepts, méthodes

Les articles scientifiques comme source d'information ?

References

Cambrosio A, Keating P, Lewison G, Mercier S, Mogoutov A., in press, Mapping the emergence and development of translational cancer research; European Journal of Cancer.

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.K., Roco, M.C., 2003. Longitudinal patent analysis for nanoscale science and engineering: country, institution and technology field. Journal of Nanoparticle Research 5, 333–363.

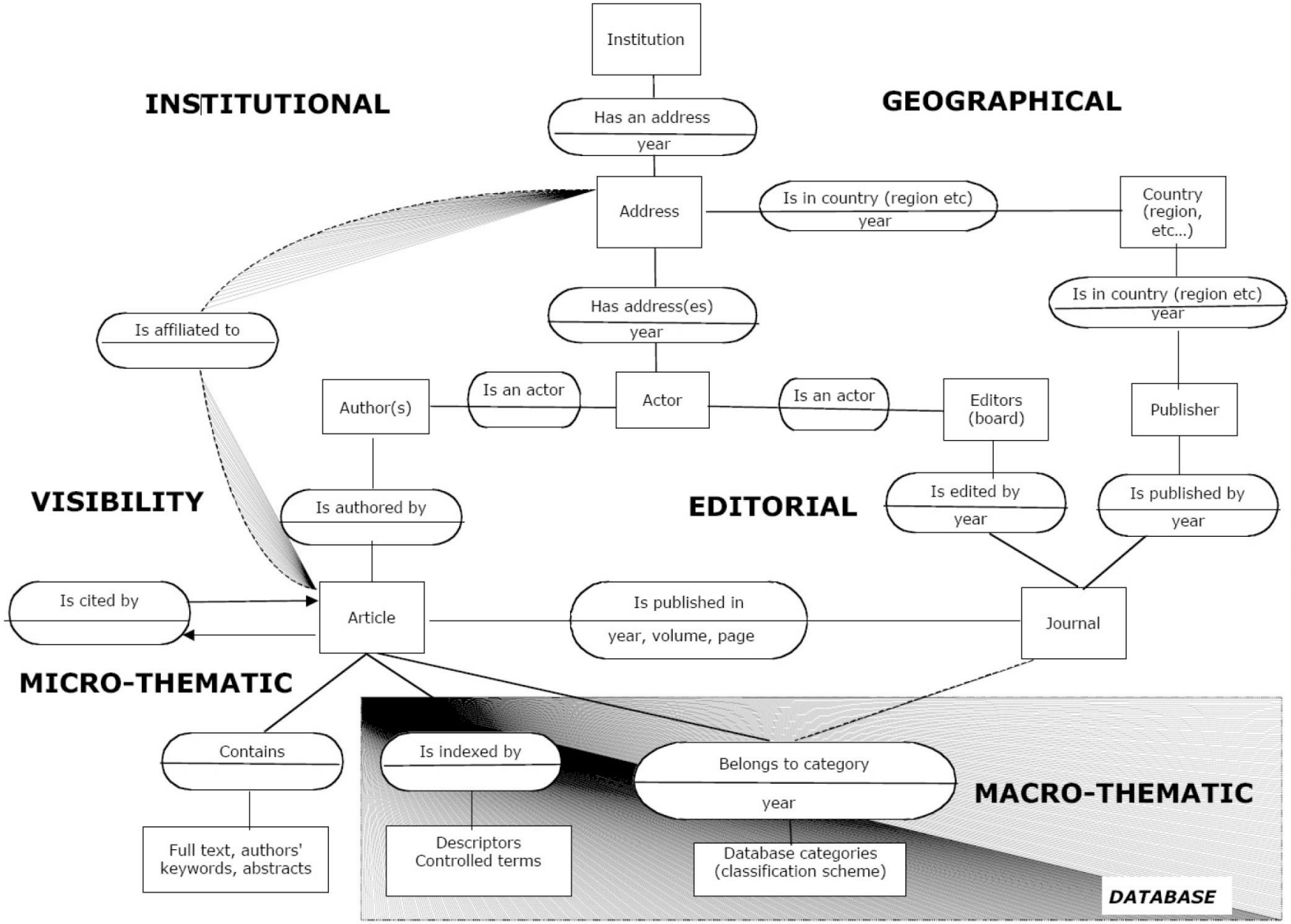
Noyons E.C.M., Buter B.K., Van Raan A.F.J., Schmoch U., Heinze T., Hinze S., Rangnow R., 2003, Mapping Excellence in Science and Technology across Europe, Nanoscience and Nanotechnology, Draft report of project EC-PPN CT-2002-0001 to the European Commission.

Sampat, B.N., 2005, Examining patent examination: An analysis of examiner and applicant generated prior art., Working Paper, Columbia University.

Zitt, M. and Bassecoulard, E., in press, "Delineating Complex Scientific Fields by A Hybrid Lexical-Citation Method: An Application to Nanosciences "Information Processing and Management".

Citations et références de l'article : sources scientifiques de l'article

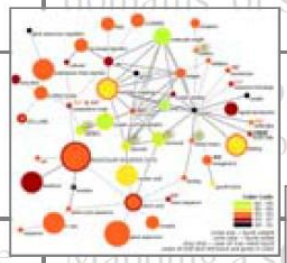
Les articles scientifiques comme source d'information ?

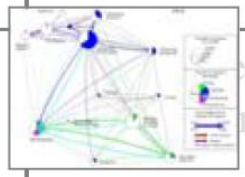
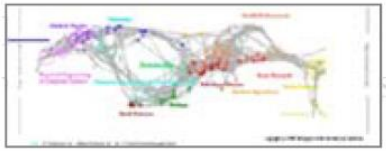


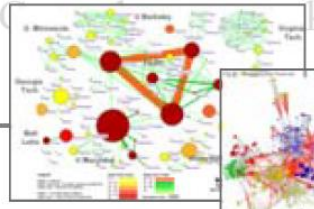
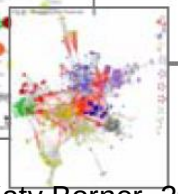
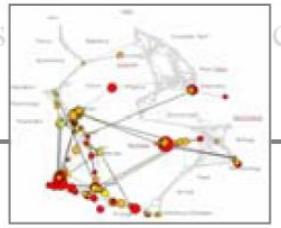
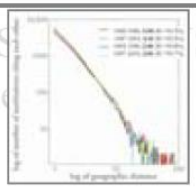
Zitt M. 2004

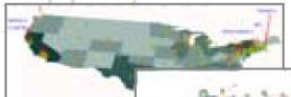
Les articles scientifiques comme source d'information ?

	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS SA, all of sci
Temporal Analysis (When)	Funding portfolio of one individual	ic bursts f PNAS	113 Years of P Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a s intellectual l	PNAS
Topical Analysis (What)	s.	research	VxOrd/Topic NIH funding
Network Analysis (With Whom?)	NSF work of one	k	NIH's cy






Atlas of science - Visualizing What We Know, Katy Borner, 2010, The MIT Press, 272 p.

Mining, Mapping, and Acceleration Science and Technology, Katy Borner, 2012, Sciences Po, paris

Les indicateurs d'activité (Callon et al., 1993)

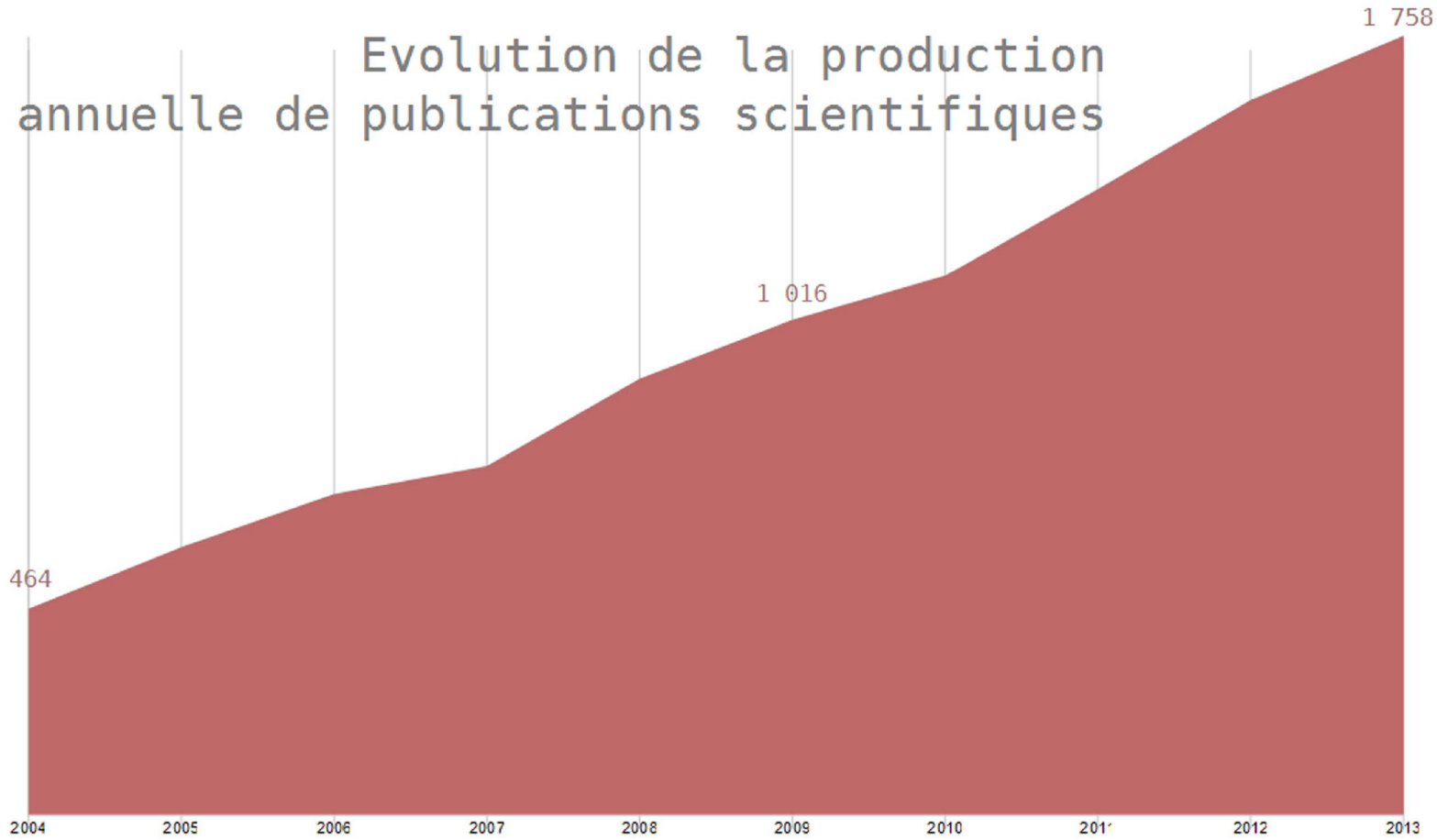
Deux grands type d'indicateurs de l'activité scientifique :

- indicateurs d'activité
- indicateurs relationnels

Les indicateurs d'activités :

- sont les plus simples
- la science est considérée comme une activité ordinaire
- il s'agit bien souvent d'un comptage des publications

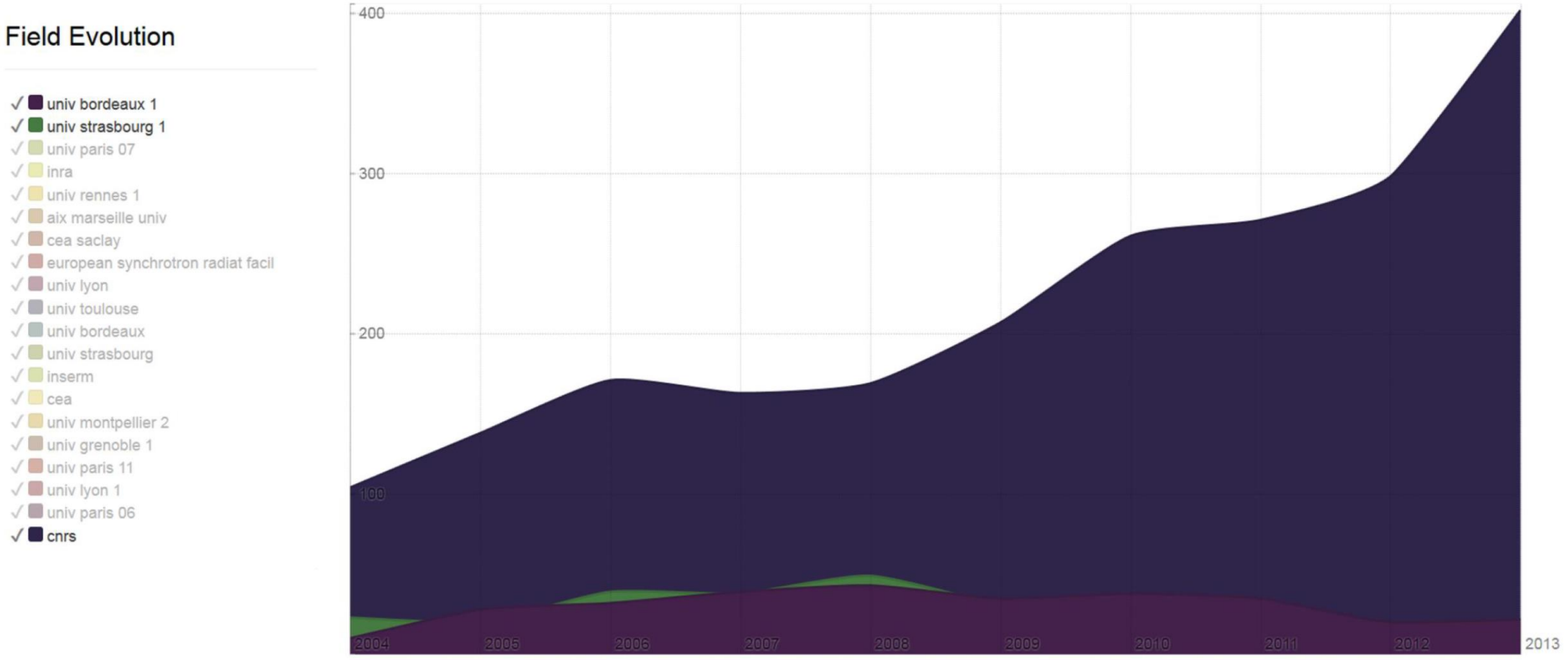
Les indicateurs d'activité (Callon et al., 1993)



Quelle est l'évolution du champ scientifique des nanobiotechnologies en France ?

Champ en forte croissance depuis les années 2004.

Les indicateurs d'activité (Callon et al., 1993)



Quelle est l'évolution de la production scientifique des principaux acteurs ?

Forte croissance, avec plusieurs profils qui semblent se dessiner, principalement entre les acteurs les plus importants (croissance forte) et ceux plus marginaux (croissance nulle) dans le développement des nanobiotechnologies en France.

Les indicateurs relationnels directs

Deux grands types d'indicateurs relationnels de l'activité scientifique :

- **directs** : relations n'entrent pas directement dans les contenus des articles (ex : adresses pour les collaborations)
- **indirects** : relations établies à partir d'une analyse du contenu des articles (ex : mots des titres, des résumés...)

Les réseaux de collaborations

Lorsqu'on étudie un ensemble d'articles, il est possible à partir des adresses des auteurs de reconstruire des réseaux de collaborations avec, par exemple :

- **Les pays** : ensemble des pays ayant collaboré avec au moins un auteur français. Cela permet dans le sujet traité de connaitre quels ont les pays avec lesquels les scientifiques français ont des relations privilégiées, et avec quelle intensité.

Les réseaux de collaborations

Top	Pays	NbPublications
1	france	45712
2	usa	1894
3	germany	1721
4	italy	1689
5	spain	1499
6	uk	826
7	japan	783
8	china	655
9	belgium	614
10	poland	588
11	russia	579

Vue en liste

Publications scientifiques dans le champ des nanotechnologies entre 1972 et 2015 dont au moins un des auteurs a une adresse en France.

Vue en matrice

Répartition des publications de 5 pays européens (Callon et al., 1993)

Tableau 5. — **Répartition des co-publications de 5 pays européens avec l'étranger** (1986)

Co-publications de	*Co-publications avec* Etats-Unis	CEE*	Japon	Reste du monde	Total
France	23,3	33,2	2,3	41,2	100
Allemagne	27,6	29,8	3,5	39,1	100
Grande-Bretagne	29,4	27,5	2,4	40,7	100
Pays-Bas	25,0	43,2	2,0	29,8	100
Italie	26,8	43,6	1,2	28,4	100
Moyenne	26,4	35,4	2,3	35,9	100

* Estimation pour la CEE, en extrapolant à partir des 5 plus grands pays.

(Source : OST.)

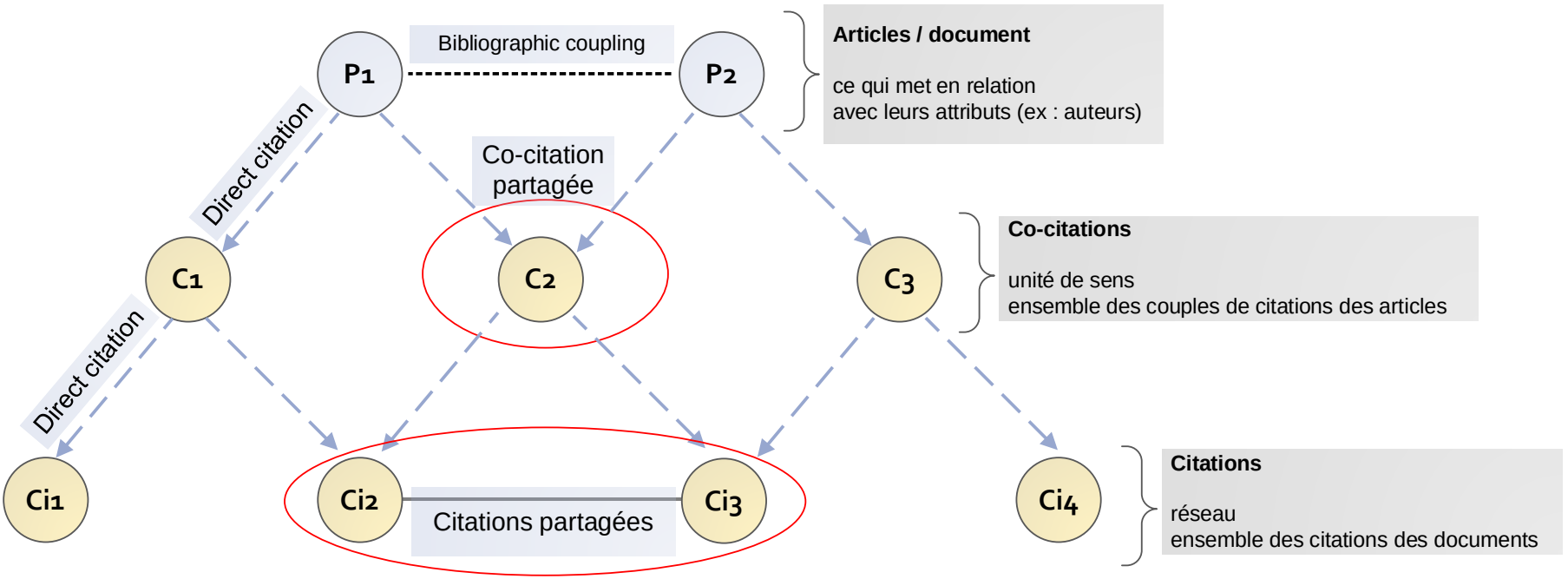
Identification des sources scientifiques

Le **réseau des citations** (des références d'un groupe d'articles) permet d'identifier quels sont les principaux travaux mobilisés et comment ils sont associés. C'est à dire les sources desquels s'inspirent les articles. Cela permet d'appréhender la visibilité par le taux de citations.

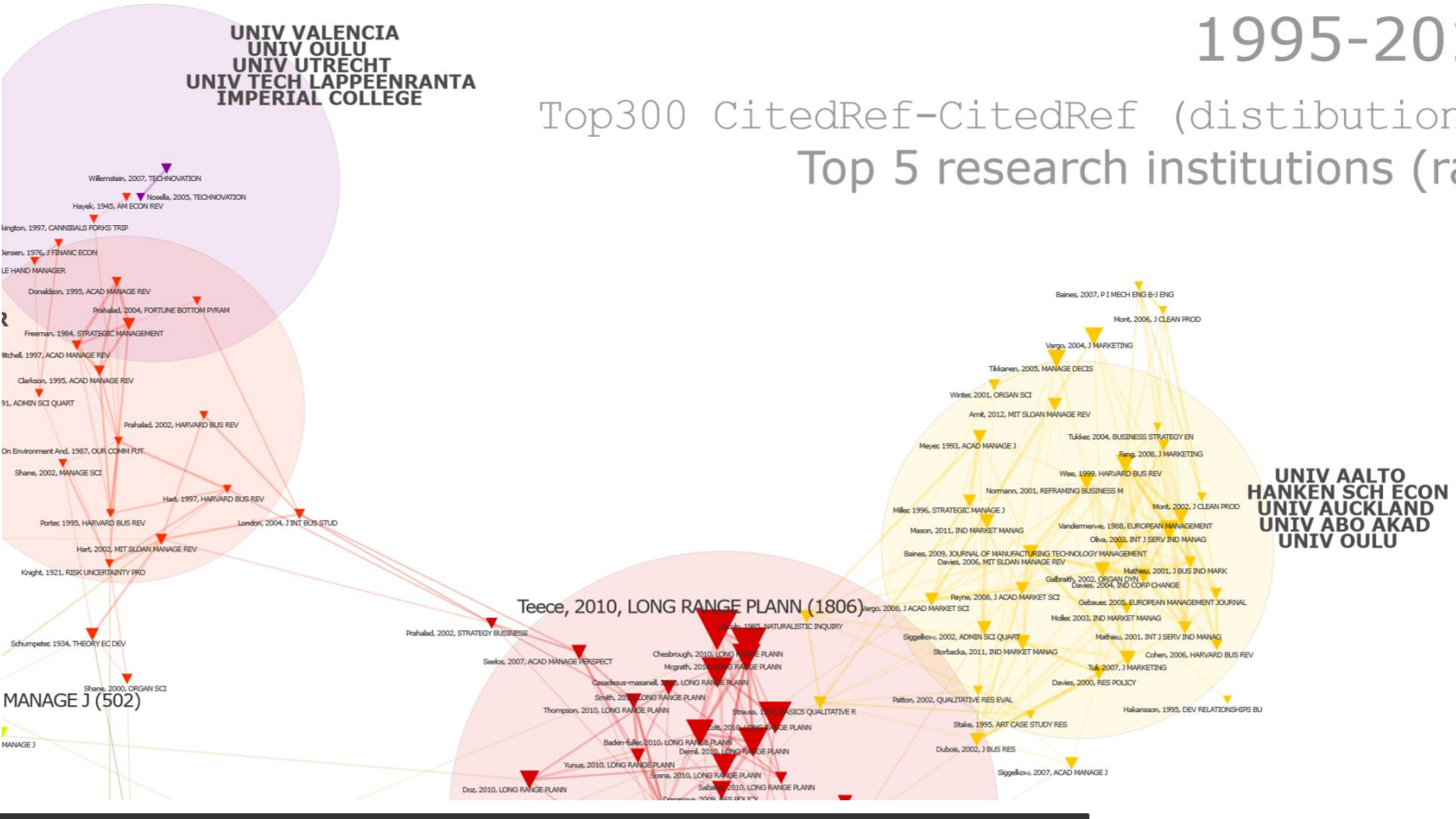
La méthode des **co-citations**, également nommée «bibliographic coupling» vise à identifier l'apparition simultanée de deux citations (couple de citations) dans l'ensemble des articles étudiés. Cette répétition dans le corpus de l'association des deux citations laisse supposer que ce couple est doté d'une signification plus précise, **plus pertinente**, que les deux citations prises indépendamment.

Identification des sources scientifiques

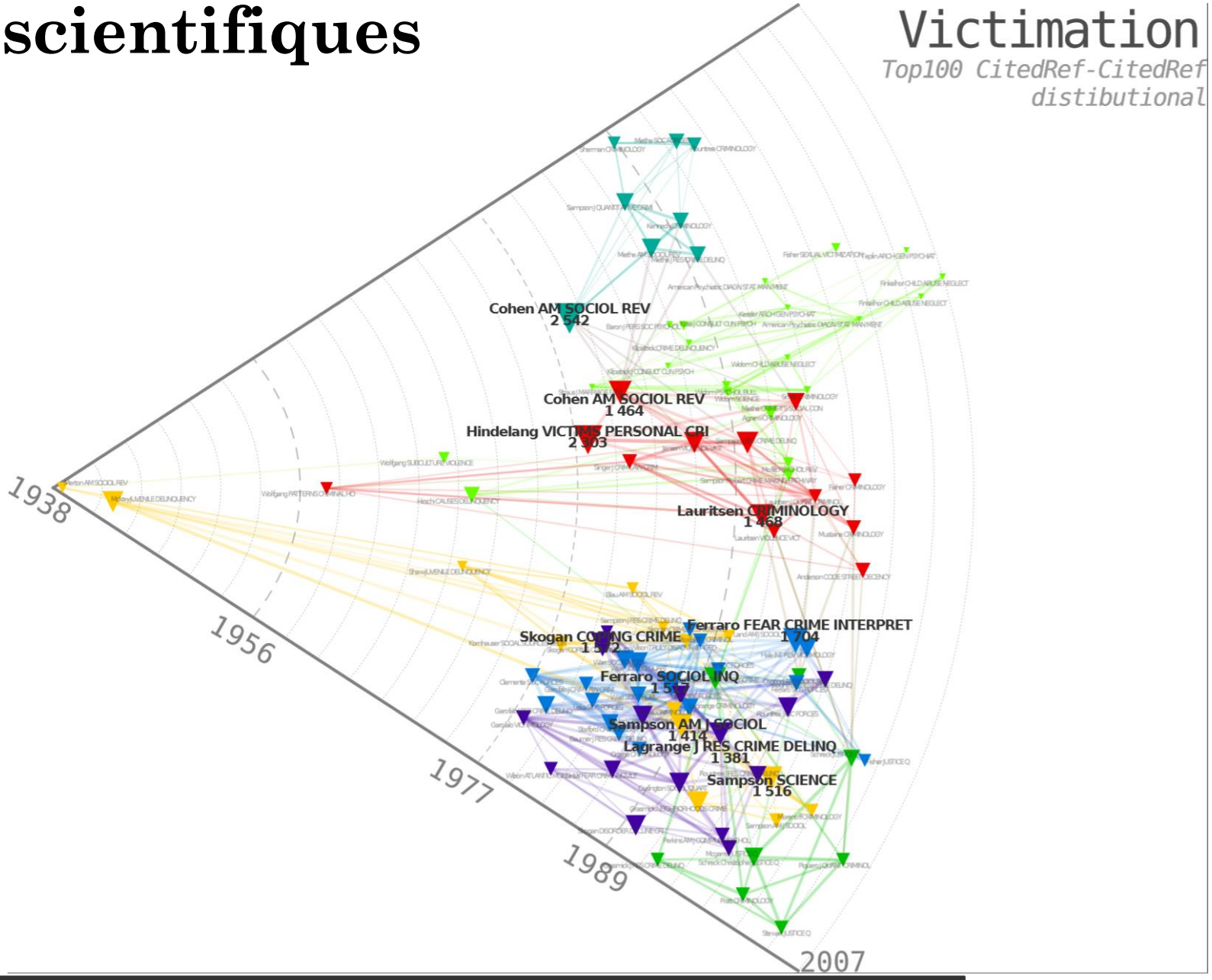
Aussi, les articles fréquemment co-cités par un groupe d'auteurs peuvent laisser supposer que ces auteurs partagent les mêmes sources scientifiques, les mêmes articles fondateurs, et peuvent témoigner d'une **communauté de chercheurs** partageant la même vision de leurs travaux.



Identification des sources scientifiques



Identification des sources scientifiques

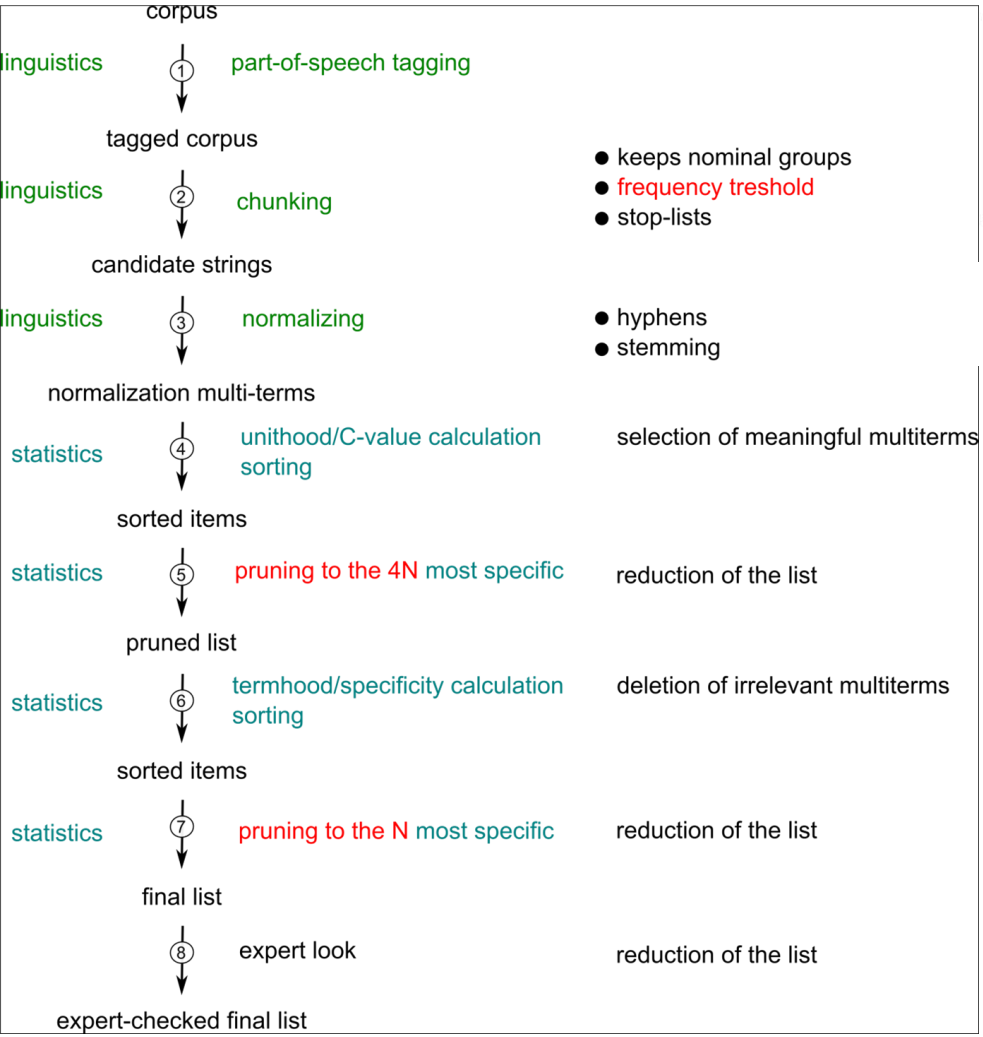


Les indicateurs relationnels indirects

indirects : relations établies à partir d'une analyse du contenu des articles (mots des titres, des résumés…)

Dans la filiation de l'analyse des co-citations, il s'agit ici d'identifier les paires de mots fréquemment répétées dans les textes des articles : permet d'appréhender la **signification des textes**.

https://docs.cortext.net/lexical-extraction



The phylogenetic position of the elephant shark (Callorhinchus milii) is particularly

DT JJ NN IN DT NN NN (NNS NN)VBZ RB

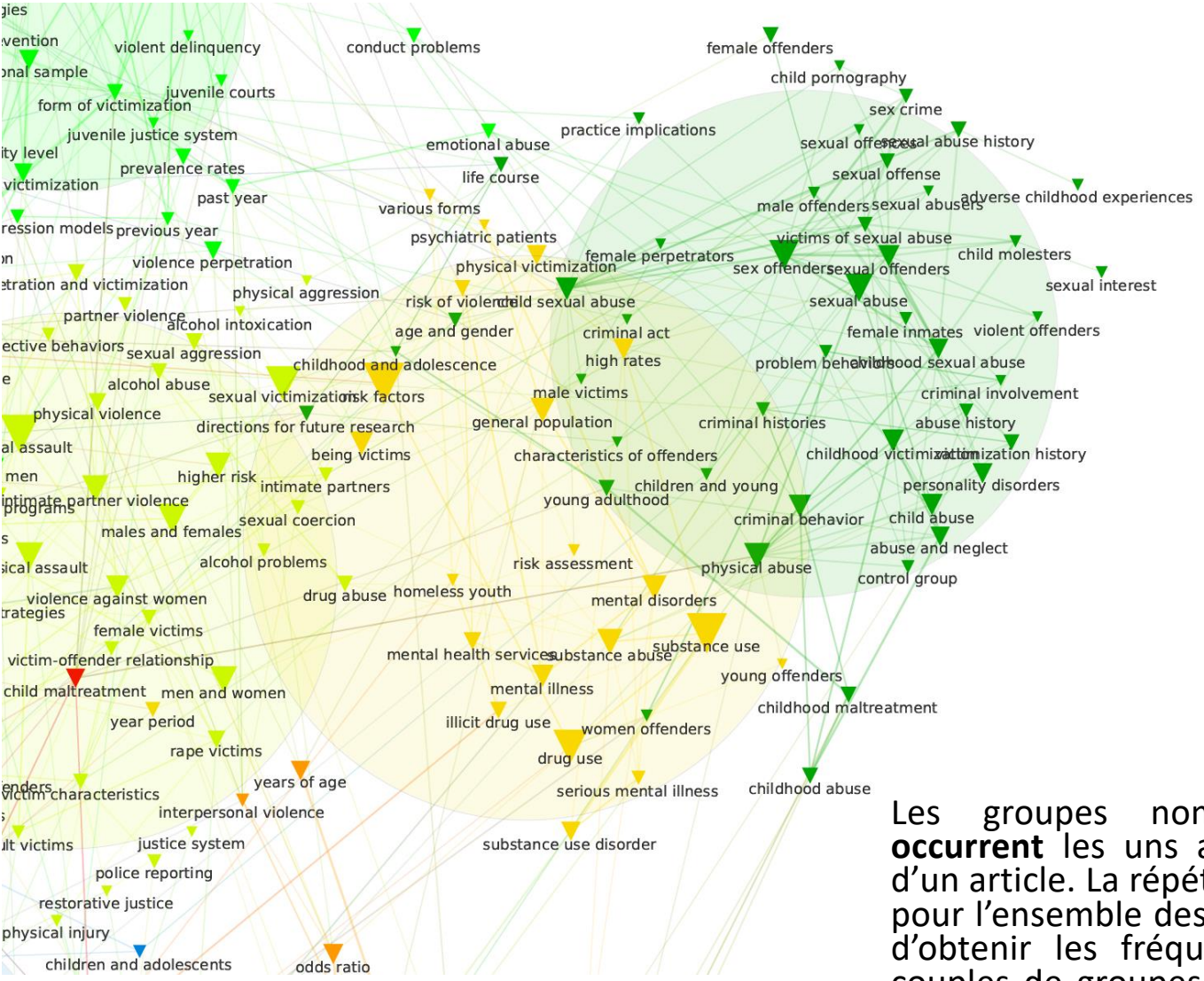
relevant to study the evolution of genes and gene regulation in vertebrates.

JJ TO VB DT NN IN NNS CC NN NN IN NNS

```
gene regulation in vertebrate -> {gene regul vertebr}
phylogenetic position of the elephant shark : {eleph phylogenet posit shark}
phylogenetic position -> {phylogenet posit}
```

stem	main form	forms	n	C-value	Specificity	Frequency
alga red	red algae	red algae|&|RED ALGAE|&|Red algae|&|red alga	2	703,3	1292,3	464
matter organ	organic matter	organic matter|&|Organic matter	2	457,1	751,0	365
chlamydomona reinhardtii	Chlamydomonas reinhardtii	Chlamydomonas reinhardtii	2	399,1	4914,1	336
higher plant	higher plants	higher plants|&|HIGHER PLANTS|&|higher plant	2	519,8	1809,2	326
acid amino	amino acids	amino acids|&|amino acid	2	281,7	1180,8	324
lactuca ulva	Ulva lactuca	Ulva lactuca	2	429,7	531,3	296

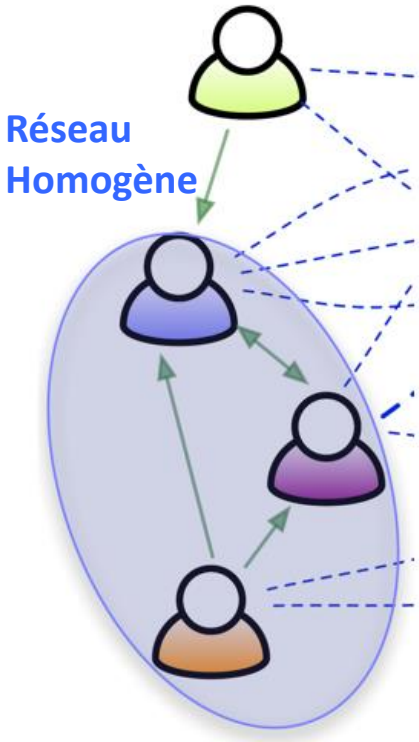
Relations sémantiques



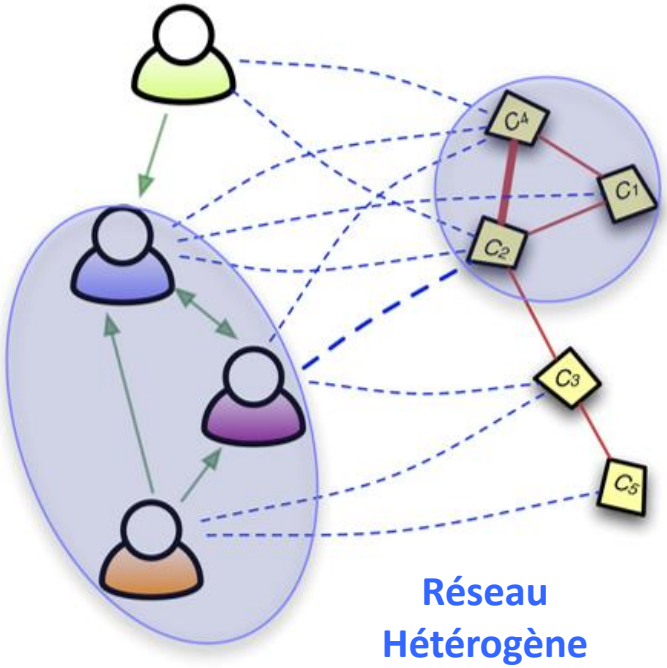
Articles sur les enquêtes portant sur la victimisation

Les groupes nominaux identifiés **co-occurrent** les uns avec les autres au sein d'un article. La répétition de cette opération pour l'ensemble des articles étudiés permet d'obtenir les fréquences d'apparition des couples de groupes nominaux et **offre une vue synthétique des principaux thèmes traités** dans l'ensemble des articles.

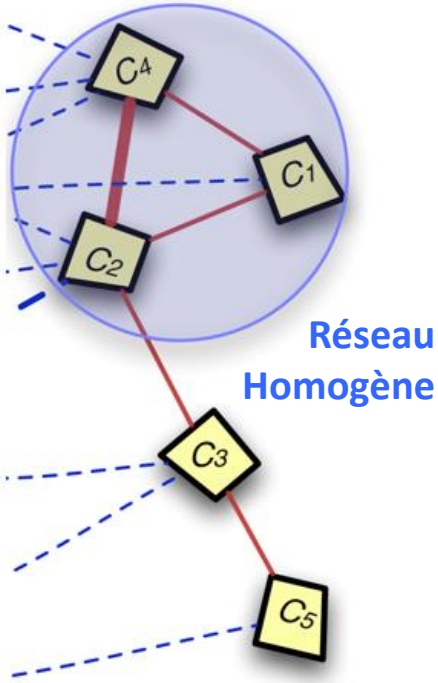
Combiner relations directes et indirectes : les liens sociaux sémantiques



Des humains ont des relations: graphe sociologique

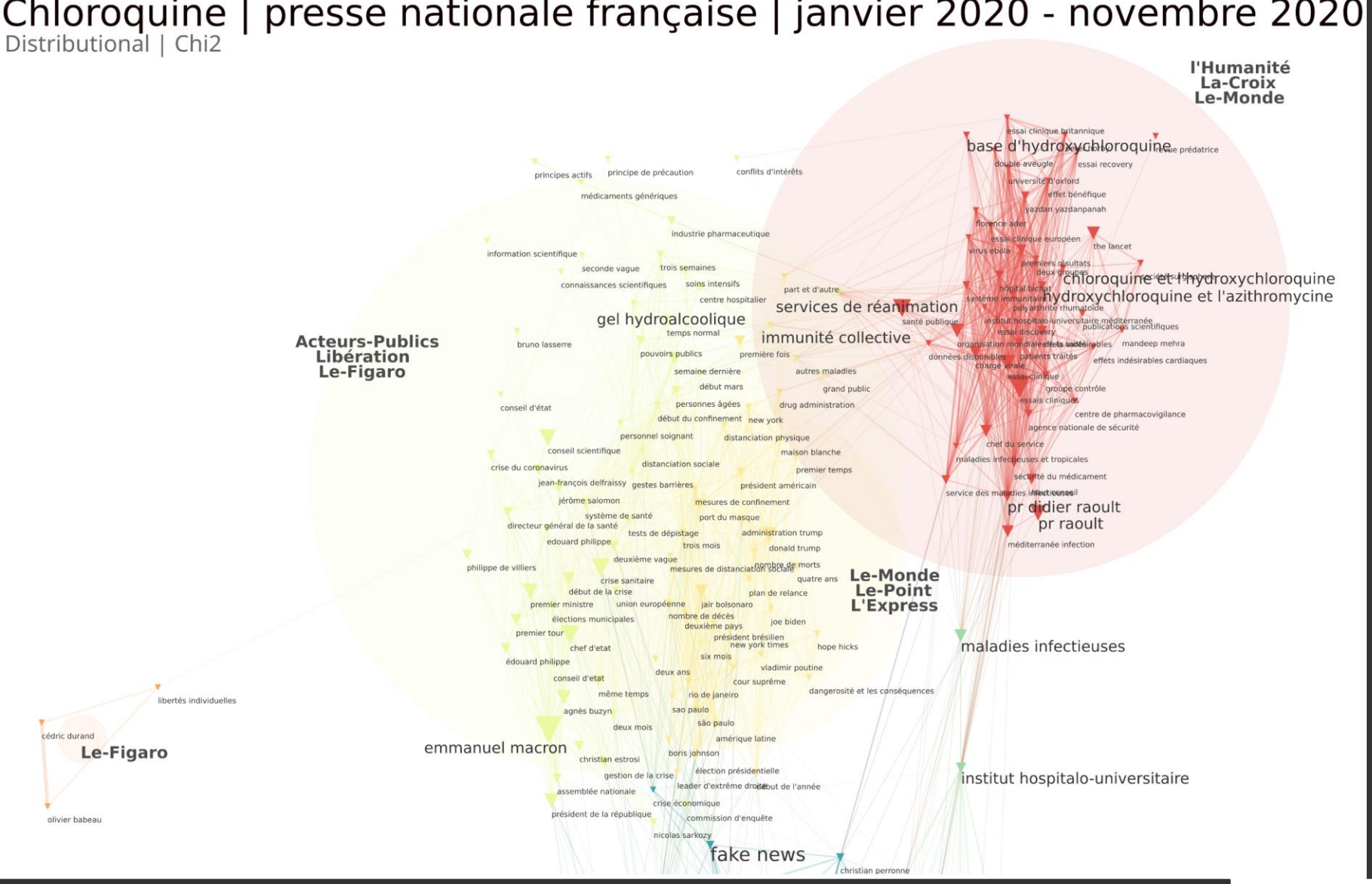


Des humains et des termes ont des relations: graphe hétérogènes (socio-sémantiques)



Des termes sont associés dans des phrases: graphe textuel

Combiner relations directes et indirectes

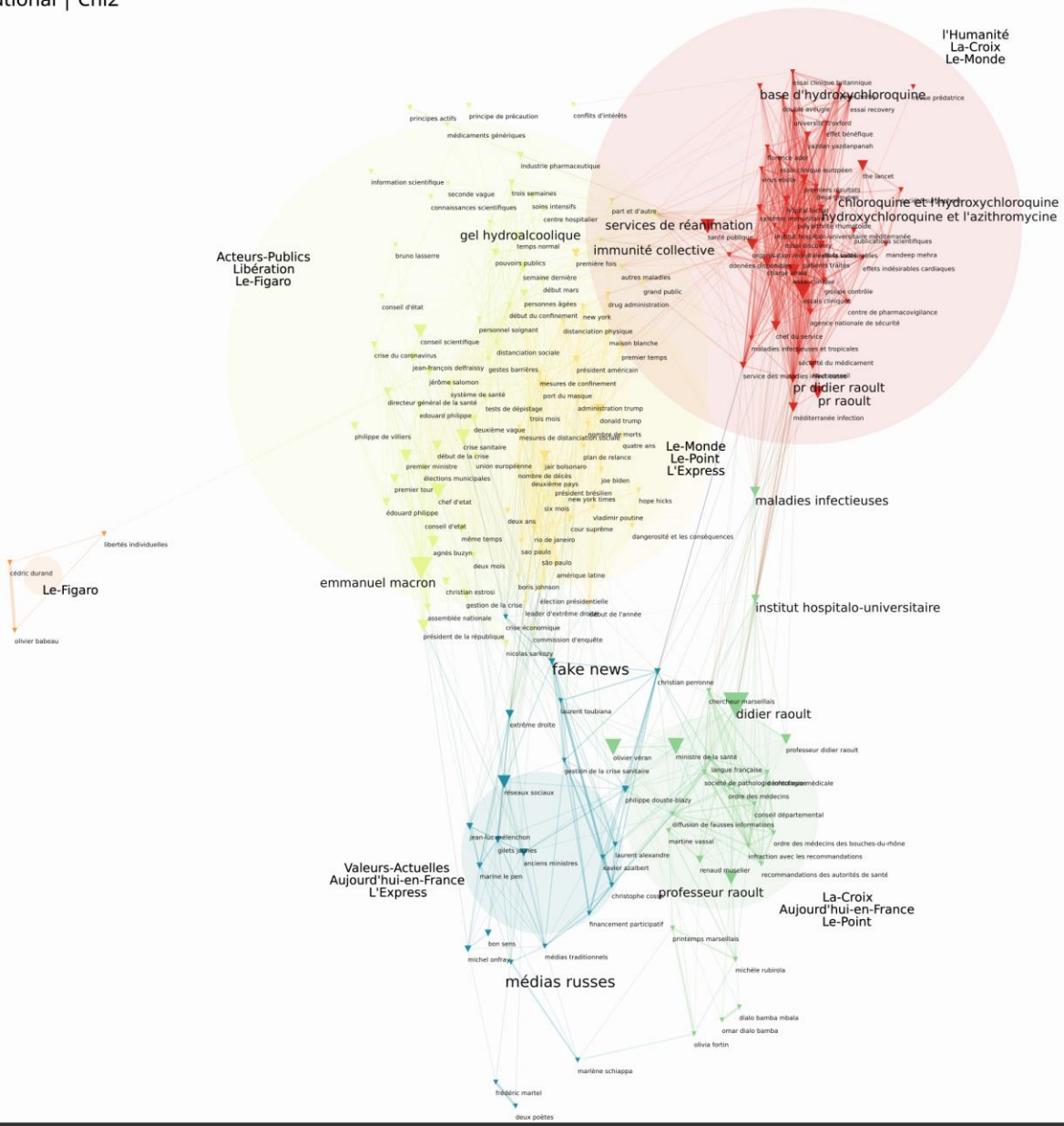


Visualisations et les trois niveaux de lecture

Lecture macroscopique

Chloroquine | French national newspapers | January 2020 - November 2020

Distributional | Chi2

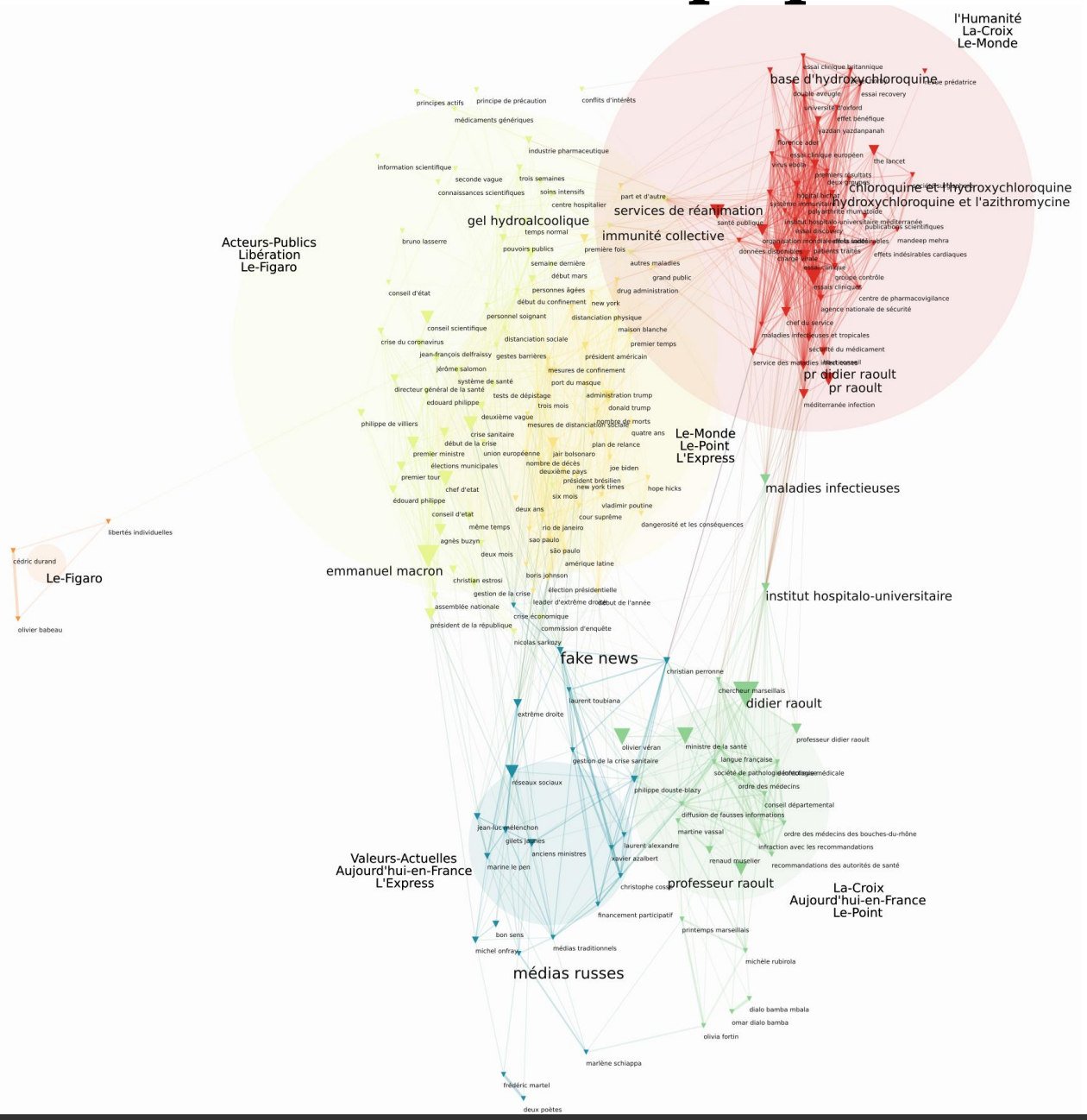


Nombre de clusters (espaces sémantiques)

Exemples de métriques

- Densité du réseau
- Tailles
- Centralisé ou distribué
- …

Lecture mésoscopique

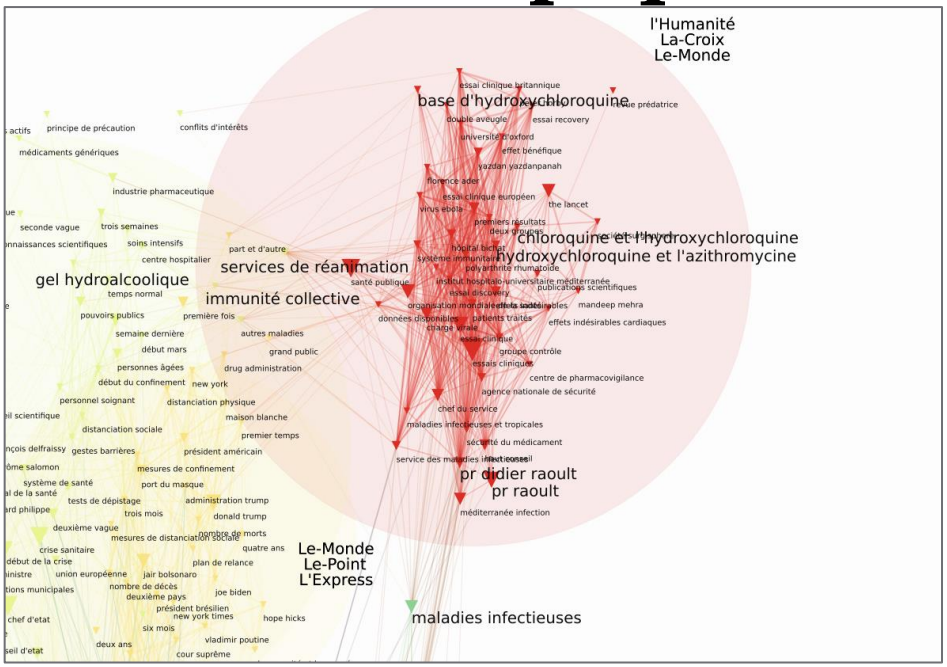


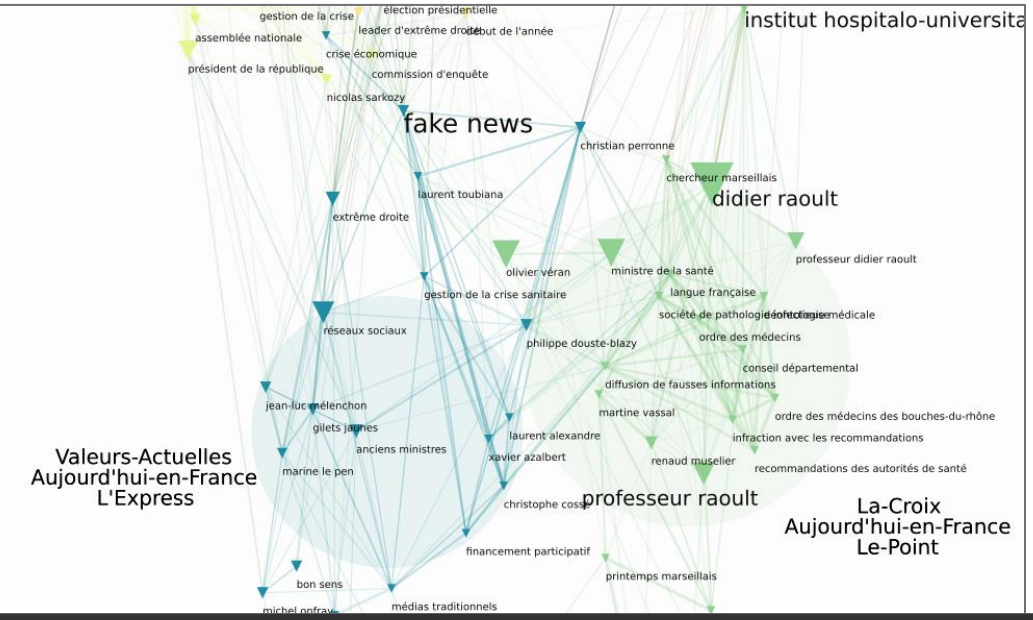
Proximité entre les clusters (espaces sémantiques interstitiels)

Exemples de métriques

- Nombre de liens ou de documents partagés

Lecture microscopique





Caractérisation des espaces sémantiques

Et des positions des nœuds

Exemples de métriques

- Centralités des nœuds
- Compositions

A vous de jouer

https://docs.cortext.net/trainings/cortext-lisis/

Et

https://managerv2.cortext.net/project/101440003210