

# CorTextT Manager : extraction d'information et analyse socio-sémantique pour les sciences humaines et sociales

CNRS TDM, Octobre 2022

Marc Barbier, Lionel VILLARD

LISIS, IFRIS, INRAe, CorTextT, ESIEE Paris

# La plateforme CORTEXT

Constituer une plateforme scientifique et technique pour soutenir un **espace de recherche sur les traces et les usages numériques de la science et de l'innovation en société.**

*“L’objectif premier de la plateforme CorText est qu’un chercheur en sciences sociales – ou un autre utilisateur – puisse venir avec une **question de recherche** et partir en **profitant des fruits d’une méthode computationnelle** adaptée à sa question.”*



avec les soutiens de



# La plateforme CorText : entre interfaces et infrastructures

## Interface humaine

Interactions avec un membre de la PF dans le cadre d'un accompagnement technique ou méthodologique

Membres  
PF

Outils

## Interface logicielle

Interactions avec les interfaces logicielles développées. En utilisant une application web, un service ou un accès aux données

Résultats  
Visualisations

*Les interfaces*

*La plateforme CorText  
ressources humaines*

*Les infrastructures*

Données

Analyses

## Matériel

Gestion, déploiement et administration des serveurs pour assurer un service constant, la sécurité des accès et des données

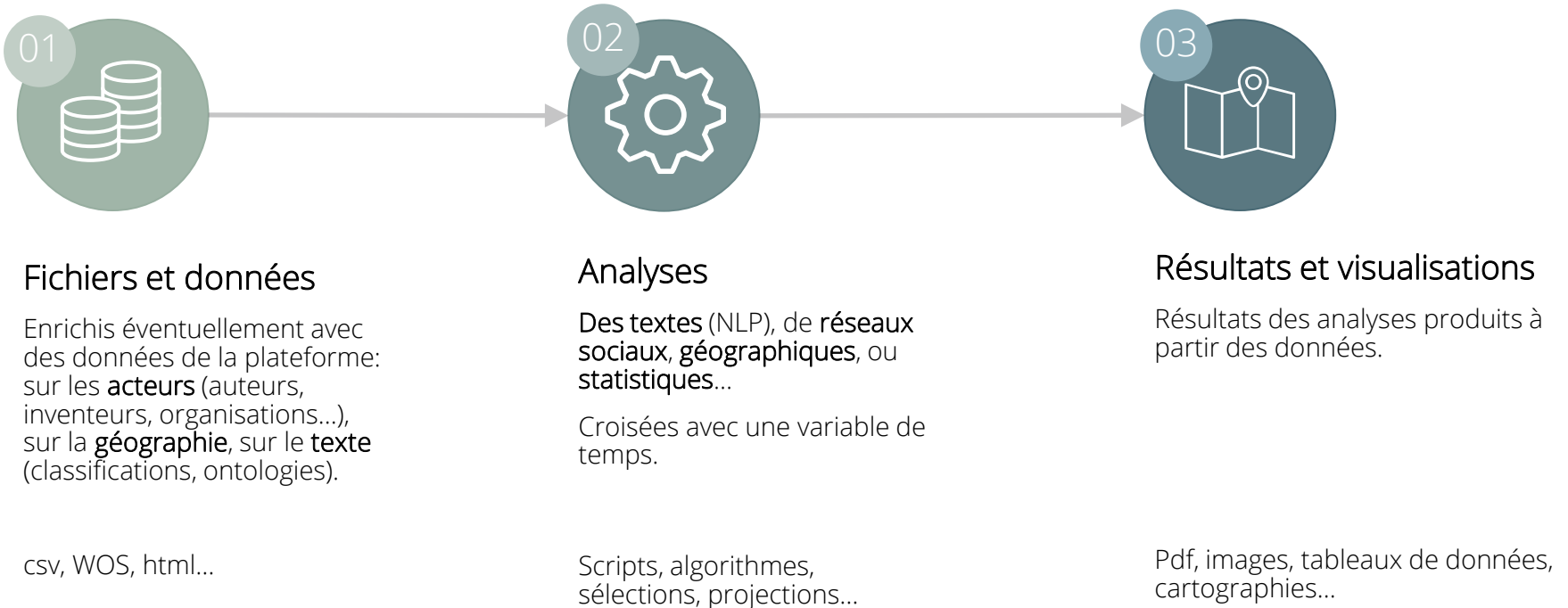
Infrastructure  
Matérielle

Infrastructure  
logicielle

## Logiciel

Développement et maintenance des logiciels (interfaces web, méthodes, en interaction avec les outils et les usagers) et des services

# Vision « usagers » : les trois étapes entre la question et l'exploitation des résultats (itératif)



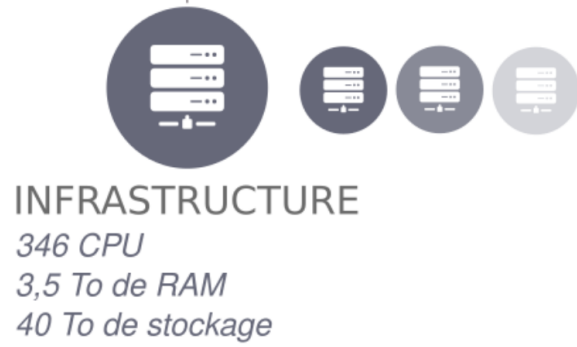
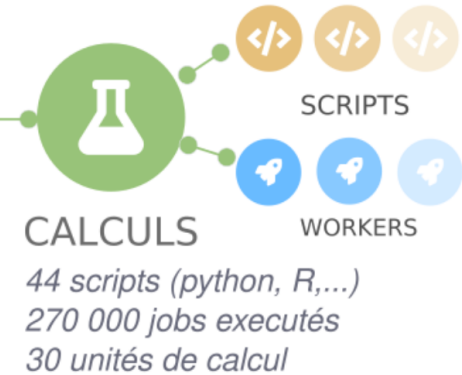
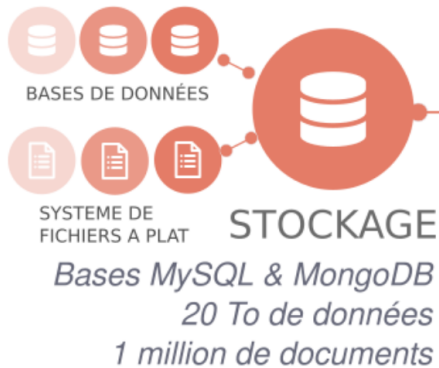
5200 utilisateurs  
105 pays

Identification  
Données brutes  
Paramètres d'analyse  
Annotations

Données analysées  
Visualisations  
Informations



**APPLICATIONS  
EXTERNES**  
*Divers projets utilisant  
l'API Cortex*





**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  
[✉](#) [🐦](#) [in](#)



**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.  
[🐦](#) [in](#)



**Bilel Benbouzid**  
Researcher, Senior lecturer at LISIS



**Alexandre Hannud Abdo**  
Post-doctorant, LISIS



**Pierre-Yves Bulot**  
IT Engineer Assistant, Cortext



**Luis-Daniel Medina**  
IT Engineer, Cortext



**Diego-Fernando Gómez Peña**  
IT Engineer, Cortext  
[in](#)



**Tatiana Andrea Sánchez Castaño**  
IT Engineer, Cortext  
[in](#)



**Joenio Marques da Costa**  
Research Software Engineer, Cortext  
[🐦](#) [in](#) [📷](#)



**Géraldine Enderli**  
Engineer specialised in the  
production, processing and analysis  
of data and survey at INRAE - LISIS,  
CorText



**Hajar Lagliil**  
Meteorological engineer/Data  
scientist, Cortext

INTERACTION CHERCHEURS,  
PROJETS, FORMATIONS,  
VALORISATION, COMMUNICATION



5 CHERCHEURS  
1 ASSISTANT INGÉNIEUR

COLLECTE DE DONNÉES,  
STOCKAGE ET CALCUL



1 INGÉNIEUR

DEVELOPPEMENT D'INTERFACES,  
VISUALISATIONS, FRONTEND



1 INGÉNIEUR

DEVELOPPEMENT DE SERVICES,  
API, BACKEND



1 INGÉNIEUR

ARCHITECTURE  
INFRASTRUCTURE, DEVOPS



1 INGÉNIEUR

METHODES D'ANALYSES,  
STATISTIQUES, SCRIPTS



1 INGÉNIEUR DE RECHERCHE  
2 INGÉNIEURS D'ETUDES

GESTION DES SERVEURS,  
SUPPORT,  
SECURITÉ



1 INGÉNIEUR SYSTÈME MI-TEMPS



**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)  
[in](#)



**Chloé Duloquin**  
Web Designer, Graphiste, Intégratrice  
web



**Jean-Philippe Cointet**  
Associate Professor, Sciences Po  
Paris, Medialab  
[in](#)



**Guillaume Orsal**  
Computer engineer, data mining, web  
development and SEO  
[🐦](#) [in](#) [📷](#)



**Cristian Martinez**  
PhD Engineer in Computer Science,  
NLP/Data Senior Consultant at  
Cogniteva  
[in](#)



**Nicolas Turenne**  
Assistant professor in data science,  
Beijing Normal University & Hong  
Kong Baptist University United  
International College  
[in](#)



**Tam Kien Duong**  
Data & design, Etalab  
[in](#)



**Nicolas Baya-Laffitte**  
STSLab, Université de Lausanne  
[in](#)



**Loïc Boudoulec**  
IT Engineer



**Bertha Brenes**  
IT Engineer, Trainee, Cortext



**Anis Arabi**  
Big data engineer  
[in](#)



**Nicolas Ricci**  
Web developer and data  
[in](#)



**Audrey Baneyx**  
Project manager Data science,  
Sciences Po Paris - medialab  
[in](#)



**Andrei Mogoutov**  
BulleScience



**Élise Tancoigne**  
Researcher, University of Geneva,  
Switzerland

# Champs disciplinaires mobilisés

- **Scientométrie**
- **Analyse des réseaux sociaux**
- Traitement automatique de la langue
- **Statistiques**
- **Cartographie géographique et analyse spatiale**
- **Visualisation de données**
- Les **méthodes en SHS**, notamment numériques mais pas seulement
- Développement logiciel



# CorText Manager en 2021

- 1313 utilisateurs actifs, générant plus de 65000 calculs;
- provenant d'environ 550 institutions (universités, entreprises, ministères, cabinets de conseil, journaliste, services de veille...) et villes



## Les publications de CorText Manager

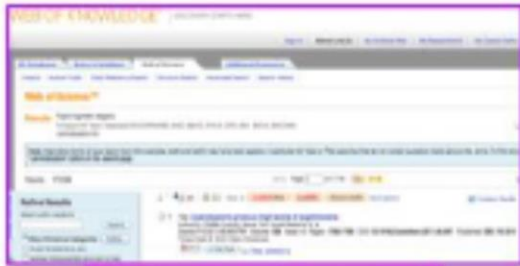
- Plus de 640 auteurs ont publié en revendiquant une utilisation de CorText Manager depuis 2016. Ces auteurs représentent moins de 10% de la communauté d'utilisateurs de CorText Manager;
- En dehors de la France : augmentations importantes en Chine, aux Philippines et au Brésil, avec la structuration de deux communautés à Wuhan et Manille.

→ <https://www.cortext.net/publications/>



# Des jeux de données multiples

## scientific productions



Web Of Science ISI

## specific databases



rare disease database

## media productions (press+web)



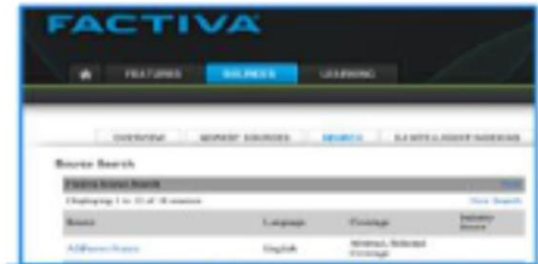
web crawler



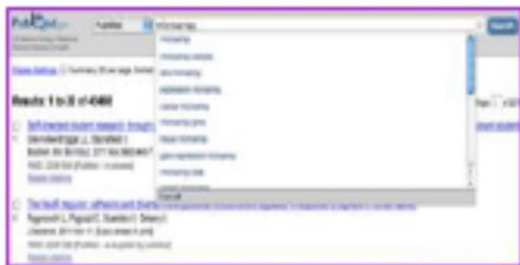
Microsoft Academic Search



projects database



Factiva, press articles archive



Medline Pubmed



clinical trials database



online forums

# Ce que CorText Manager produit

## Deux principaux types d'indicateurs

- Les indicateurs simples de **statistique descriptive** : stocks, rangs, fréquences
- **Indicateurs relationnels** (réseaux), avec deux sous types :
  - **Natifs**: qui s'appuient sur des variables directement accessibles dans les données initiales, par exemple dans des métadonnées. Ces indicateurs ne rentrent donc pas dans le contenu des documents.  
*Par exemple : collaboration entre les auteurs...*
  - **Ajoutés** : qui s'appuient les résultats d'une analyse des contenus des documents. Les réseaux qui en sont issus sont donc dérivés d'un calcul effectué.  
*Par exemples : réseau des cooccurrences des mots des textes des documents, ou encore le réseau de collaboration entre les aires métropolitaines...*

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



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Research Policy 36 (2007) 893–903



Journal

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

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

Andrei Mogoutov<sup>a,\*</sup>, Bernard Kahane<sup>b,c,1</sup>

Auteurs : collaboration scientifique

<sup>a</sup> AGUIDEL, 68 Bld de Port Royal, 75005 Paris, France

<sup>b</sup> 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

<sup>c</sup> 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 nanostrig 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, portability 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

© 2007 Elsevier B.V. All rights reserved.

Keywords: Datamining; Nanotechnology; Emergent technologies

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

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

## 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”.

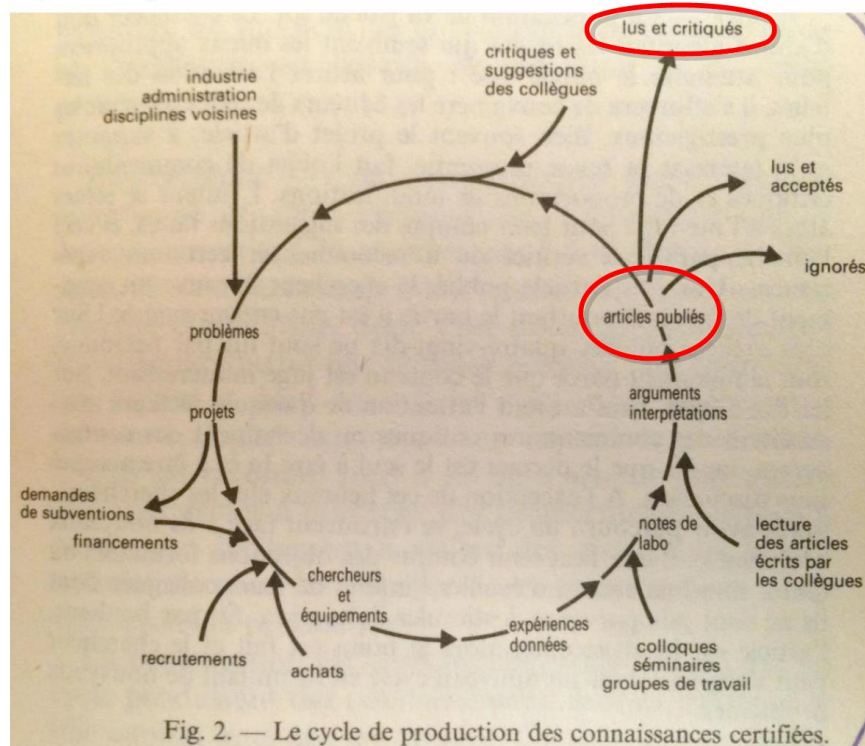
*Citations et références de l'article : sources scientifiques de l'article*

# Activité de recherche et traces numériques

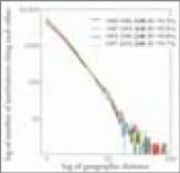


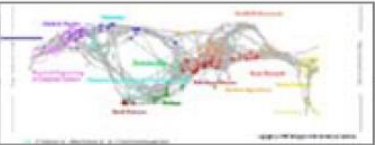
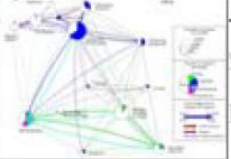


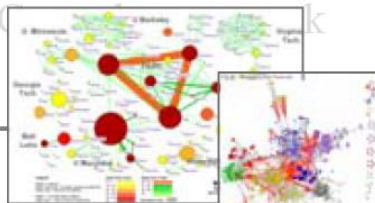
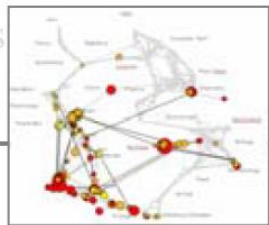
A ce titre, un **article scientifique** est considéré comme un indicateur important de la production de la recherche scientifique (mais pas le seul).

Les « **connaissances certifiées** » sont des connaissances qui ont été soumises à la critique des collègues et qui ont résisté 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).



# Dimensions d'analyse et échelles

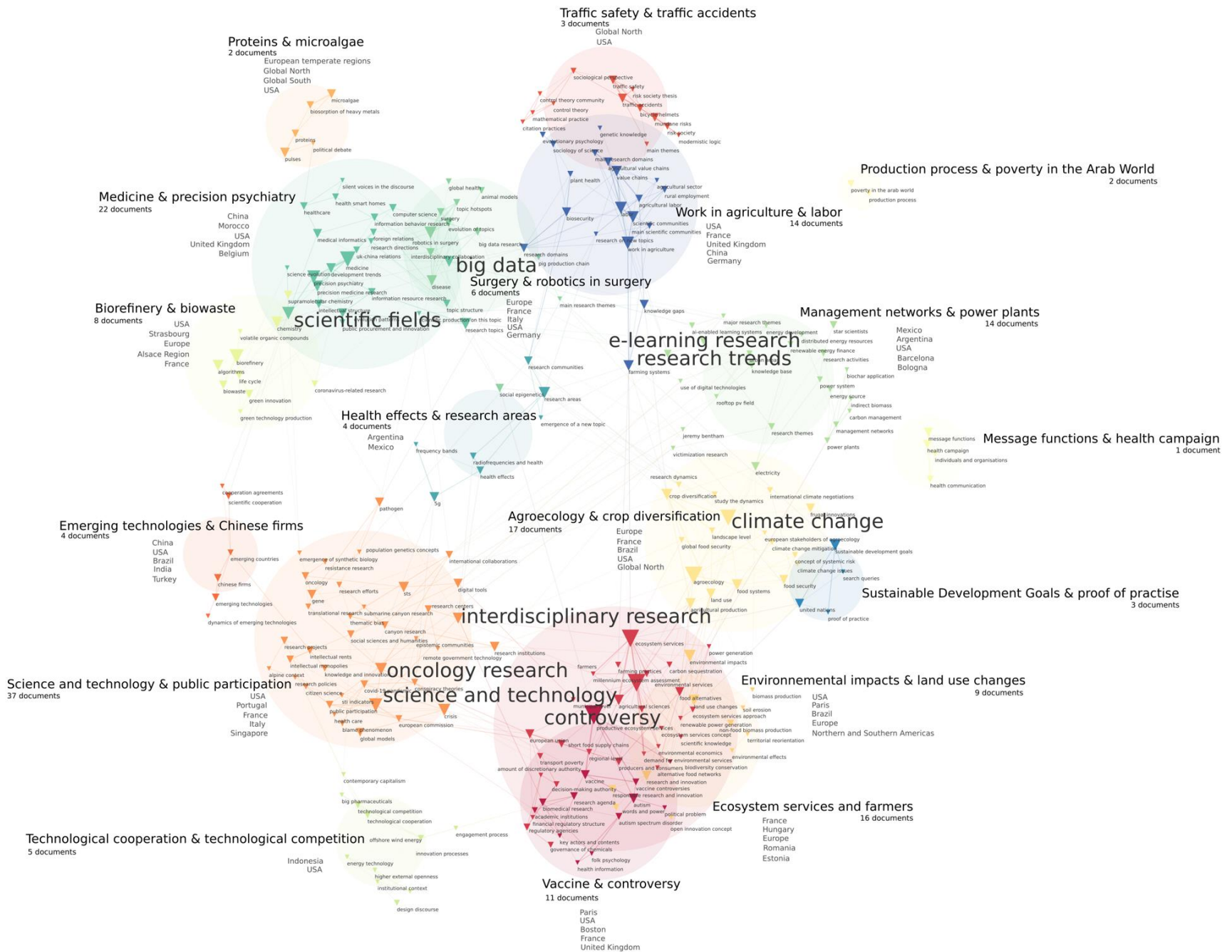
	<i>Micro/Individual (1-100 records)</i>	<i>Meso/Local (101-10,000 records)</i>	<i>Macro/Global (10,000 &lt; records)</i>
<b>Statistical Analysis/Profiling</b>	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS... SA, all of sci... 
<b>Temporal Analysis (When)</b>	Funding portfolio of one individual	...ic bursts of PNAS	113 Years of P Research 
<b>Geospatial Analysis (Where)</b>	Career trajectory of one individual	...mapping a s... intellectual l...	PNAS 
<b>Topical Analysis (What)</b>			VxOrd/Topic r NIH funding 
<b>Network Analysis (With Whom?)</b>	NSI... 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 axes de travail fréquent

- **Emergence et cartographie** de domaine de recherche et d'innovation (bioénergie, nanotechnologie, biodiversité)
- **Controverses** et « hot topics » (pesticides, biologie de synthèse, sécurité alimentaire mondiale)
- Analyse du **web et des média-sociaux** pour caractériser les phénomènes de la Science et de l'Innovation en Société
  
- Modes d'usages :
  - **Cartographie socio-sémantique et relationnelle** des productions de la recherche (publications, brevets, projets) dans une production académique
  - **Appui à la démarche qualitative** : utiliser la carte comme un support efficace pour faire parler les acteurs;
  - **Revue de la littérature** : certains auteurs utilisent CorText Manager pour circonscrire et situer leurs travaux (parcours doctoral, dans un papier).

→ <https://docs.cortext.net/trainings/cortext-metsem-2022/04-exemples-papiers/>



**Proteins & microalgae**

2 documents  
European temperate regions  
Global North  
Global South  
USA

**Traffic safety & traffic accidents**

3 documents  
Global North  
USA

**Production process & poverty in the Arab World**

2 documents  
poverty in the arab world  
production process

**Medicine & precision psychiatry**

22 documents  
China  
Morocco  
USA  
United Kingdom  
Belgium

**Work in agriculture & labor**

14 documents  
USA  
France  
United Kingdom  
China  
Germany

**Biorefinery & biowaste**

8 documents  
USA  
Strasbourg  
Europe  
Alsace Region  
France

**big data**

6 documents  
Europe  
France  
Italy  
USA  
Germany

**Surgery & robotics in surgery**

6 documents  
Europe  
France  
Italy  
USA  
Germany

**Management networks & power plants**

14 documents  
Mexico  
Argentina  
USA  
Barcelona  
Bologna

**scientific fields**

coronavirus-related research  
green technology production

**e-learning research trends**

use of digital technologies  
knowledge base

**Health effects & research areas**

4 documents  
Argentina  
Mexico  
frequency bands  
radiofrequencies and health  
health effects

**Agroecology & crop diversification**

17 documents  
Europe  
France  
Brazil  
USA  
Global North

**climate change**

17 documents  
European stakeholders of agroecology  
climate change mitigation  
sustainable development goals  
concept of systemic risk  
climate change issues  
search queries

**Message functions & health campaign**

1 document  
message functions  
health campaign  
individuals and organisations  
health communication

**Emerging technologies & Chinese firms**

4 documents  
China  
USA  
Brazil  
India  
Turkey

**interdisciplinary research**

research projects  
intellectual rents  
knowledge and innovation  
citizen science  
public participation  
health care  
biomimetic phenomenon  
global models

**Sustainable Development Goals & proof of practise**

3 documents  
proof of practice

**Science and technology & public participation**

37 documents  
USA  
Portugal  
France  
Italy  
Singapore

**oncology research**

translational research  
submarine canyon research  
research centers  
social sciences and humanities  
epistemic communities  
remote government technology  
research institutions

**Environmental impacts & land use changes**

9 documents  
USA  
Paris  
Brazil  
Europe  
Northern and Southern Americas

**science and technology controversy**

17 documents  
European Union  
libert food supply chains  
transport poverty  
amount of discretionary autonomy  
vaccine  
decision-making authority  
response  
research agenda  
biomedical research  
academic institutions  
financial regulatory structure  
regulatory agencies  
key actors and contents  
governance of algorithms  
folk psychology  
health information

**Ecosystem services and farmers**

16 documents  
France  
Hungary  
Europe  
Romania  
Estonia

**Technological cooperation & technological competition**

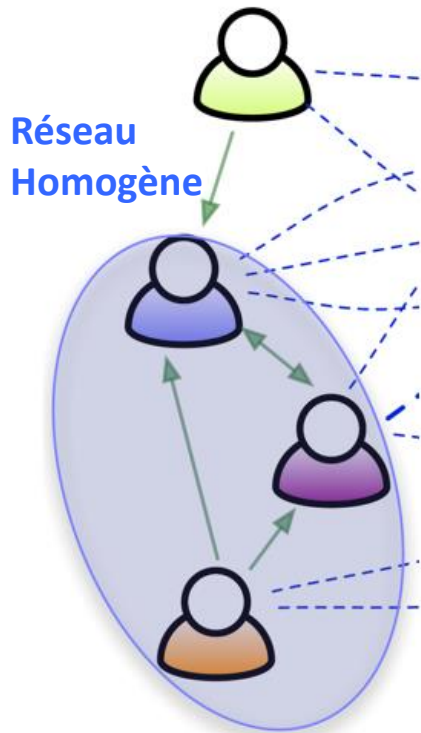
5 documents  
Indonesia  
USA  
energy technology  
higher external openness  
institutional context  
design discourse

**Vaccine & controversy**

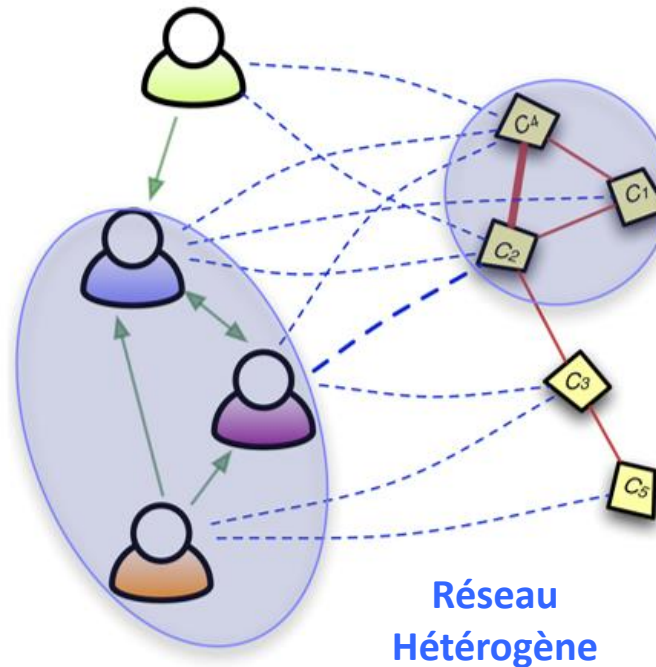
11 documents  
Paris  
USA  
Boston  
France  
United Kingdom



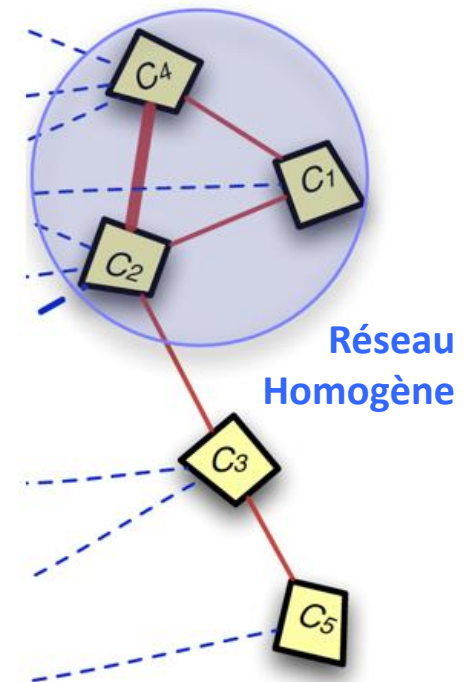
# Combiner relations directes et indirectes : réseau hétérogène



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

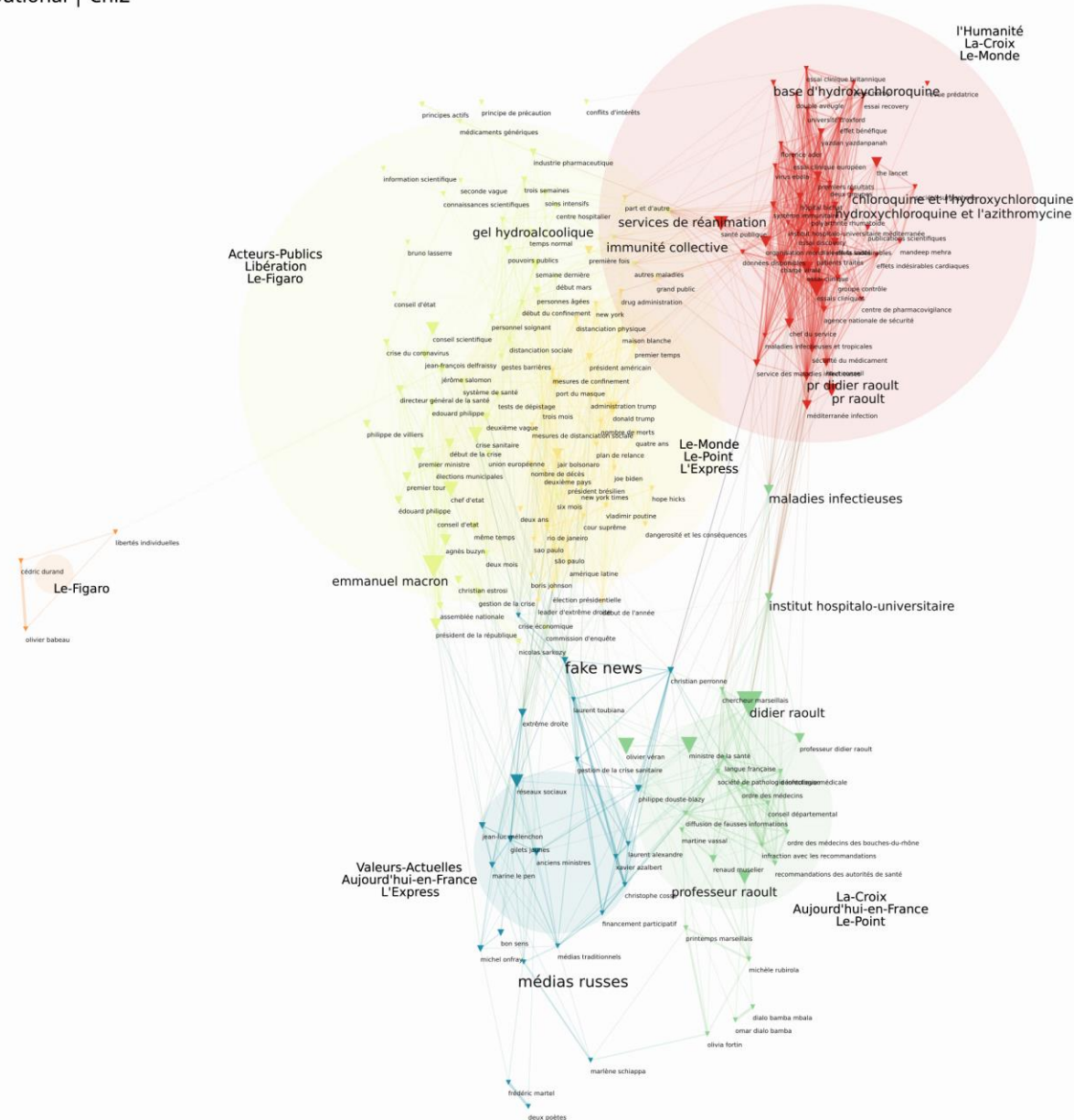
# Les mesures de proximités disponibles

proximity measures	type of network	normalisation	special properties
raw	interaction network (e.g. social network)	no	-
$\chi^2$	homogeneous & heterogeneous	yes	normalization tend to create links toward higher degree nodes
MI	homogeneous & heterogeneous	yes	Inspired from information theory
Cramer	homogeneous & heterogeneous	yes	-
cosine	homogeneous network (eg. semantic)	yes	Classical measure (originating from scientometrics)
distributional	homogeneous network (eg. semantic)	yes	very robust measure (coming from computational linguistics)
cosine_het	affiliation network (eg. users sharing the same hashtags )	yes	two fields are required but the final network is homogeneous
dot_product_het	affiliation network (eg. users sharing the same hashtags )	no	two fields are required but the final network is homogeneous

# Visualisations et les trois niveaux de lecture

# Lecture macroscopique

Chloroquine | French national newspapers | January 2020 - November 2020  
Distributional | Chi2

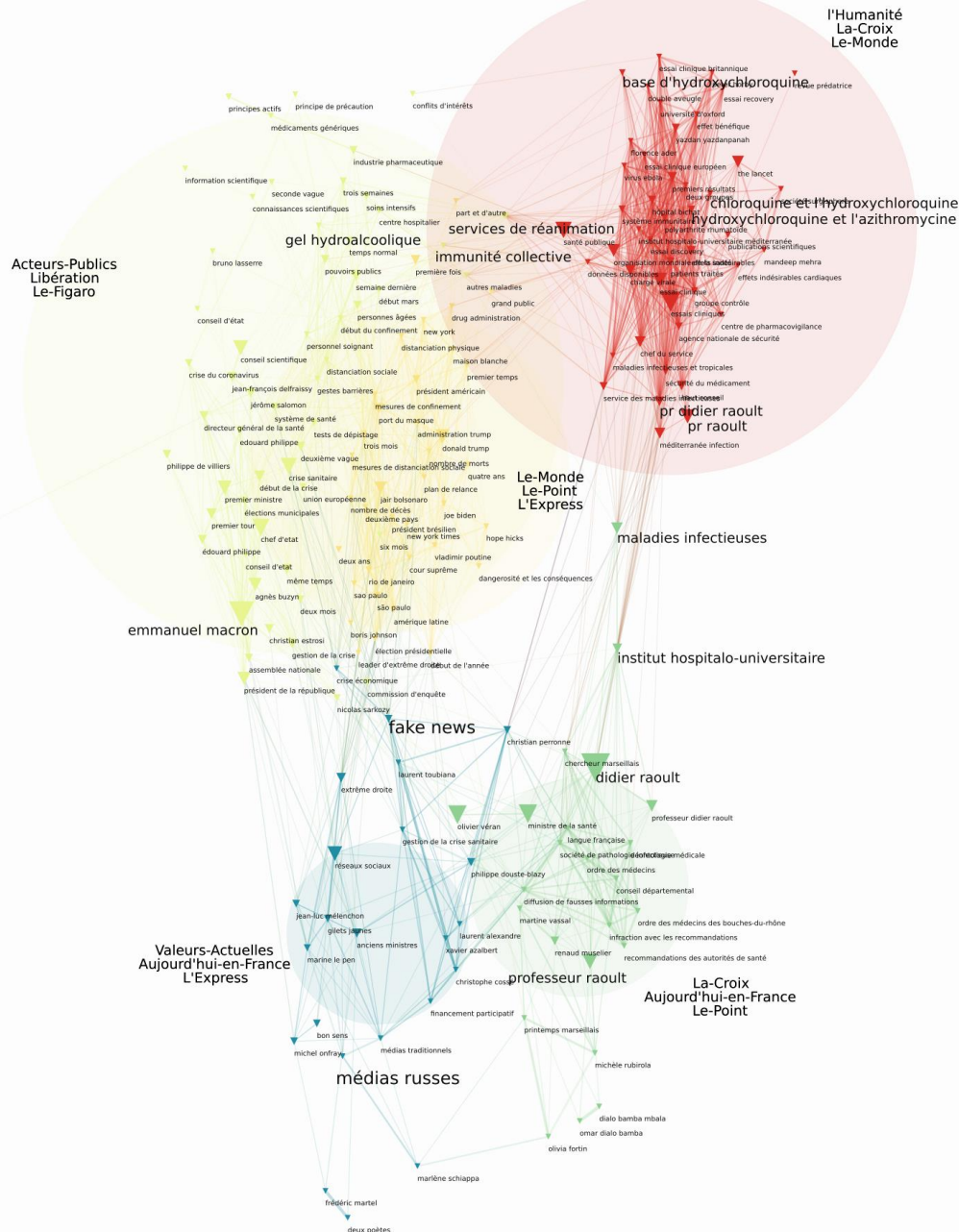


Nombre de clusters  
(espaces sémantiques)

Exemples de métriques

- Nombre de clusters
- Densité du réseau
- Réseau centralisé ou distribué
- ...

# Lecture mésoscopique

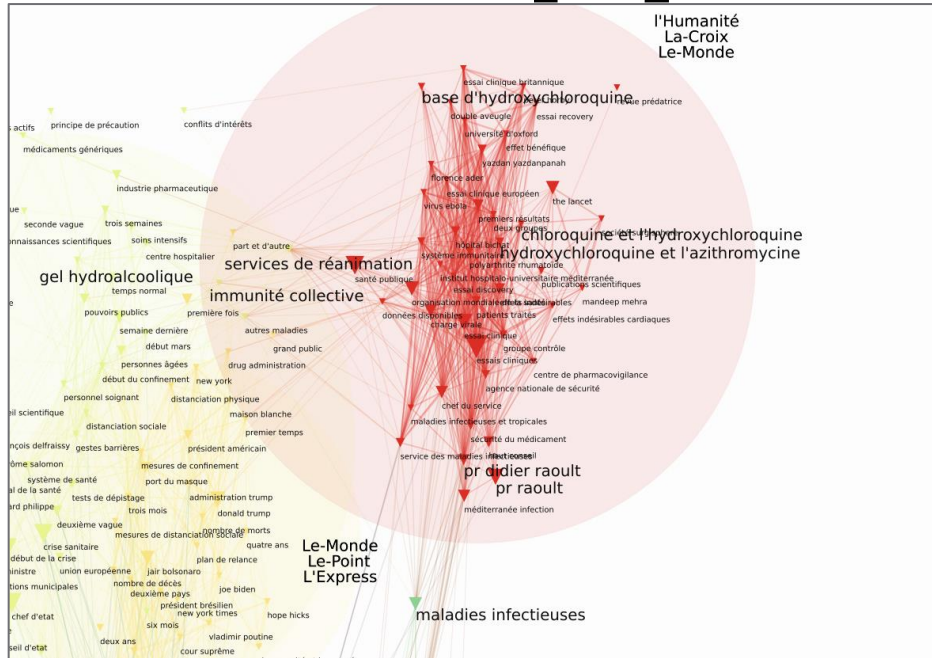


Proximité entre les clusters (espaces interstitiels) sémantiques

## Exemples de métriques

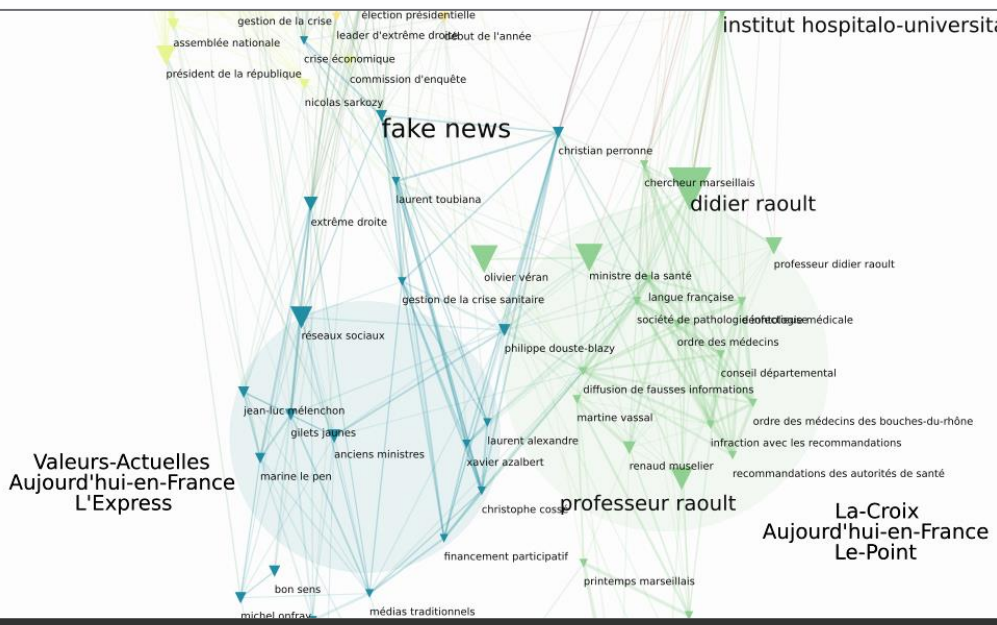
- Distances entre les clusters
- Nombre de liens ou de documents partagés
- Tailles et importances relatives des clusters
- Densités intra-clusters

# Lecture microscopique



Interprétation locale de la composition des clusters

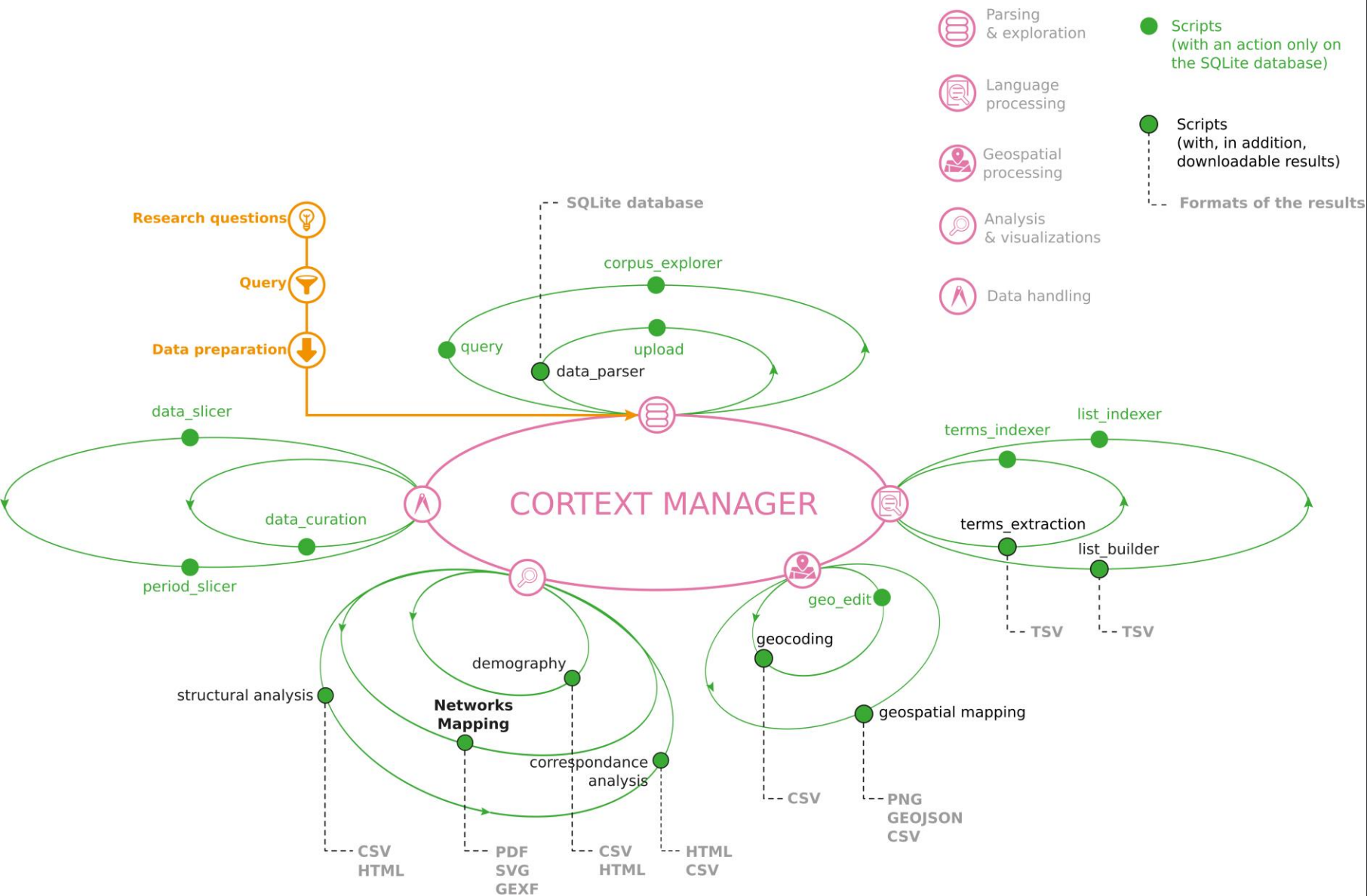
Et des positions des nœuds



Exemples de métriques

- Centralités des nœuds
- Compositions des clusters
- Chevauchements entre les clusters

# CorText Manager galaxy



# Aller plus loin

 @CorText\_team

→ <https://docs.cortext.net/trainings/cortext-cnrs-tdm-2022/>

Et

→ <https://managerv2.cortext.net/project/161250003210>