

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

CorTextT Plateform, Janvier 2024

Lionel VILLARD

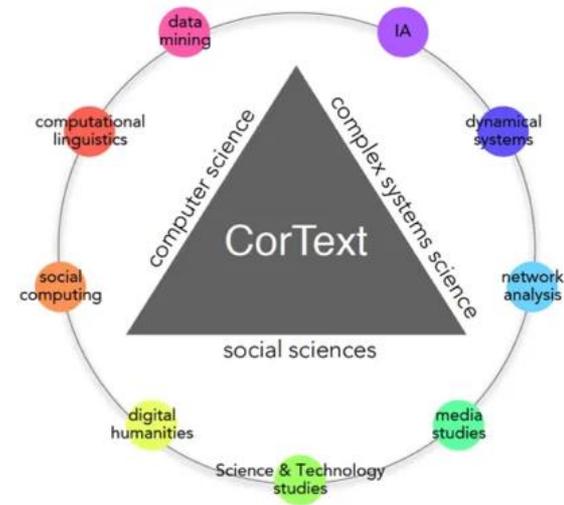
LISIS, IFRIS, INRAe, ESIEE Paris, CorTextT

# La plateforme CORTEXT

*“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.”*

*Réponse, une plateforme qui conjugue :*

- *Sciences Informatiques,*
- *Sciences de la donnée,*
- *et Sciences Humaine et Sociales.*



# 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

5200 utilisateurs  
 105 pays

Identification  
 Données brutes  
 Paramètres d'analyse  
 Annotations

Données analysées  
 Visualisations  
 Informations

WEB  
 CORTEXT MANAGER  
 SERVICES WEB  
 API "Cortext As A Service"  
 OUTILS DE VISUALISATION & EXPLORATION



APPLICATIONS  
 EXTERNES  
*Divers projets utilisant  
 l'API Cortext*

BASES DE DONNÉES  
 SYSTEME DE FICHIERS A PLAT  
**STOCKAGE**  
*Bases MySQL & MongoDB  
 20 To de données  
 1 million de documents*

SCRIPTS  
 WORKERS  
**CALCULS**  
*44 scripts (python, R,...)  
 270 000 jobs executés  
 30 unités de calcul*

MONITORING  
*Statuts des services  
 Santé de l'infrastructure  
 Etude des usages  
 Statistiques*

INFRASTRUCTURE  
 346 CPU  
 3,5 To de RAM  
 40 To de stockage



**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](#)

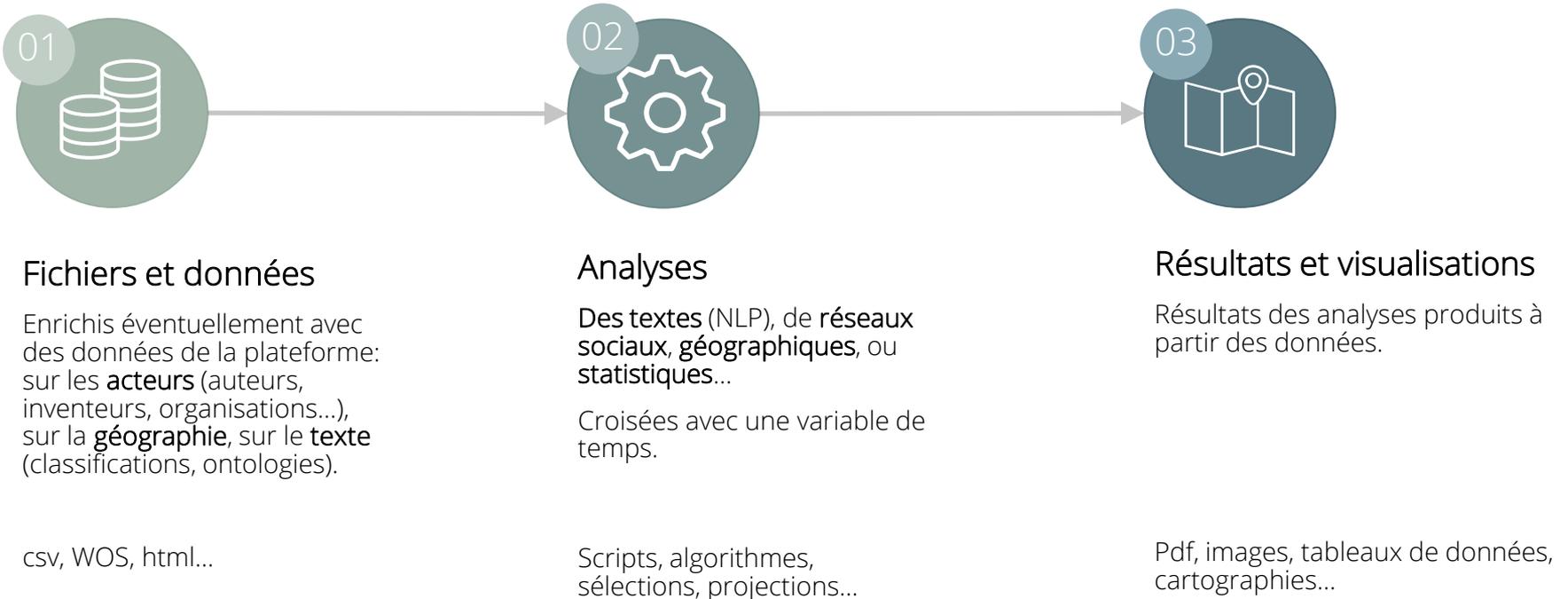


**Andrei Mogoutov**  
BulleScience



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

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



# 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



# A l'origine de l'application web CorText Manager

Né en 2009, **construites au contact de recherches en sciences sociales** et soutenu par le développement d'une infrastructure technique et de méthodes de calcul.

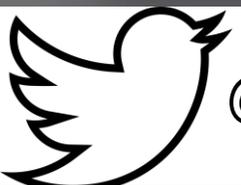
- Soutenu à l'origine par l'institut français INRA
- Rejoint par LISIS, et H2020 RISIS



Deux versions majeures de l'application web CorText Manager : v1 (2011) et v2 (2016)

- Conçue à l'origine pour d'alléger les **chercheurs en sciences sociales** du besoin en calcul intensif requis pour exécuter certains algorithmes, sans nécessiter de compétences en programmation
- **Combiner** différentes dimensions d'analyse **au sein de la même interface**
- **Collaborative**
- La **découverte peut être inattendue** : très permissive dans le traitement des types de données et des paramètres. Tous les algorithmes et méthodes sont "exécutables" sur toutes les variables.

The screenshot displays the CorText Manager web interface. At the top, there are navigation buttons: 'upload file' (pink), 'start script' (green), and 'start discussion' (yellow). The main area features a complex network graph with nodes of various sizes and colors (purple, green, yellow, blue, red) connected by lines. Each node is labeled with a research topic, such as 'SCIENTIFIC LITERATURE', 'NATURAL LANGUAGE PROCESSING KNOWLEDGE BASE SOCIAL SCIENCES', 'CLIMATE CHANGE CLIMATE NEGOTIATIONS NATIONS FRAMEWORK CONVENTION', 'ONCOLOGY TRANSLATIONAL RESEARCH CO-CITATION ANALYSIS', and 'ANIMAL MODELS'. The interface also includes a sidebar on the left with sections for 'Your analyses' (listing several 'Network Mapping' tasks), 'datasets' (showing 'cortext-blitz-v06'), and 'participants' (listing names like Lionel, Tatiana Sánchez, and Diego Fernando Gómez Peña). On the right, there is an 'Information' section with a progress indicator, a 'Cite CorText Manager' section with citation options (APA, BibTex, BibLatex) and a sample citation, and a 'Housekeeping' section with trash and download icons.



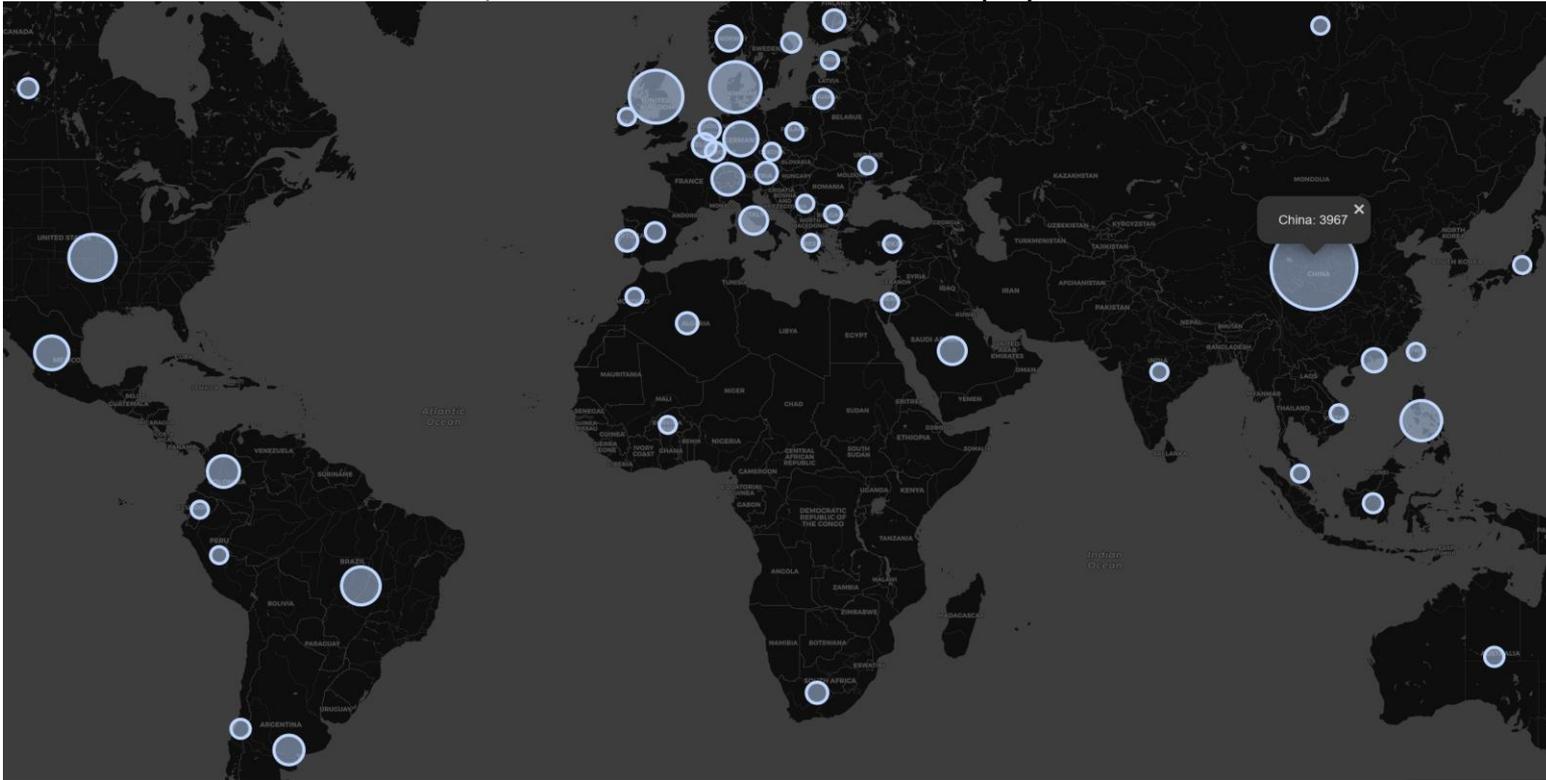
@CorTexT\_team

<https://docs.cortext.net/>



# CorText Manager en 2023

- 1190 utilisateurs actifs, générant plus de 45 000 calculs;
- provenant de 492 institutions (universités, entreprises, ministères, cabinets de conseil, journaliste, services de veille...) et villes, localisées dans 50 pays différents.



## Les publications de CorText Manager

- Plus de 1 000 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/>

# Modes d'usage identifiés

## Quatre modes d'usage dans la recherche et la formation

### 1/ Analyse **quantitative** dans une production académique

- Cartographie **socio-sémantique et relationnelle** de productions, notamment celles de la recherche (e.g. publications, brevets, projets)...
  - dont analyse **bibliométrique** (e.g. environnement en sciences regionales);
  - et analyse **scientométrique** (e.g. Scientometric review permafrost research);
- ...puis de la **presse et des media sociaux** (e.g. robots vs algorithmes).
- Étude du positionnement relatif et de la **spécialisation d'acteurs** (e.g. entreprise, université) et d'espaces géographiques (e.g. aire urbaine ou rurale, région, pays) (e.g. geographical dynamics R and D networks-robotics).

### 2/ Appui à la **démarche qualitative**

- Utiliser la carte comme un **support efficace pour alimenter le travail d'enquête** de terrain (e.g. dialogue avec les acteurs).
- Utilisation de **CorText Manager sur des produits des techniques d'enquêtes** mobilisées
- Analyse de texte pour la capture **des représentations portées par des acteurs et organisations** (e.g. divinfood food chains actors for using agrobiodiversity);
- **Analyse des réseaux sociaux** (SNA) mettant en lumière des structures organisationnelles et sociales présentes dans les situations observées (e.g. european quality assurance initiatives local gastronomy small scale farmers).

3/ **Revue de la littérature** : utilisation de CorText Manager afin de circonscrire et situer des travaux (e.g. dans un travail doctoral, dans un article scientifique).

4/ **Contextes pédagogiques** : traitement automatique des langues et analyse des textes, étude de controverses, étude des dynamiques sociales et politiques sur les médias sociaux (e.g. tendances linguistiques, mobilisations sociales, discriminations systématiques), Social Network Analysis, veille technologique et scientifique [...]. Pour l'année académique 2023-2023, onze cours ont utilisé CorText Manager dans leurs dispositifs pédagogiques (e.g. Sciences Po Paris, Aalborg University, De La Salle University, Université de Bordeaux, MINES Paris - PSL, Université Lyon 1).

## Éclairages stratégiques et rapports d'expertise pour l'aide à la décision

- **Éclairage politique ou stratégique** : note stratégique à destination du politique, veille, intelligence économique, de la conduite de la R&D, politique de la recherche, suivi de "trending topics"...
- **Rapport d'expertise** : contexte de controverse, fabrique de la norme, état de l'art...

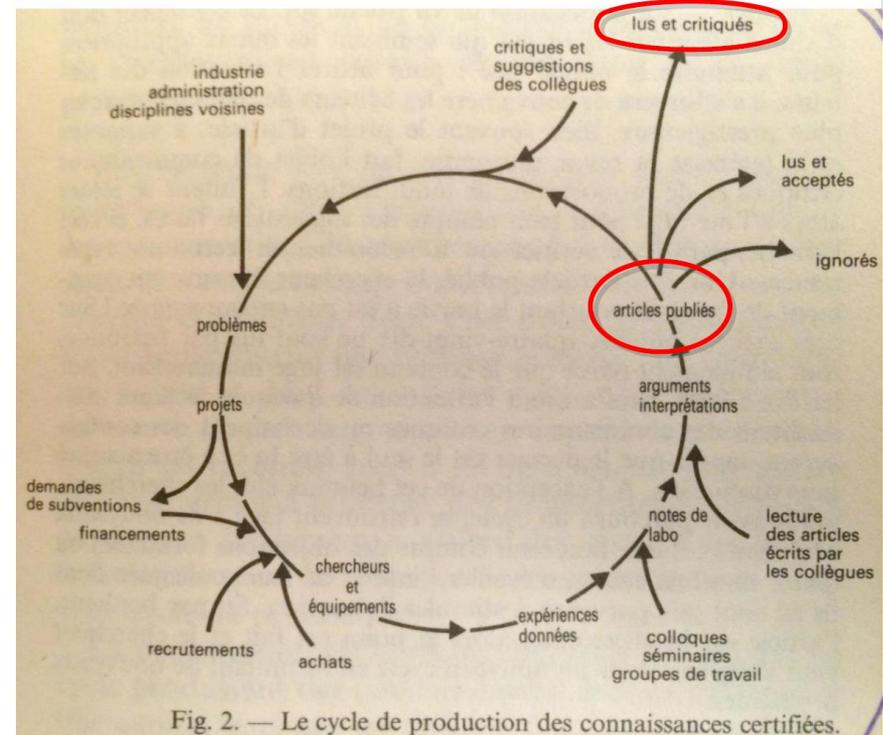
→ <https://docs.cortext.net/trainings/cortext-lisis-2024/04-exemples-papiers/>

# 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).



# 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*

# Des jeux de données multiples

JOB NAME

Data Parsing ▶ chloro-sci-2014-2020-v02.zip

job label (optional)

SCRIPT PARAMETERS

Source

Type of Data

dataset  cortext db

Corpus Format

isi

Ignore entries with

yes

isi  
ris (scopus)  
ris (standard)  
nbib  
istex  
factiva  
europresse  
guardian  
robust csv  
txt  
json  
json (multiline)  
docx  
pdf  
xls

## Scientific production

- Web of Science / ISI
- Scopus
- PubMed / nbib
- Istex : full text archive
- RIS : standard

## Press / newspapers

- Factiva
- Europress
- Guardian

## Social media

- Twitter : json

## Generic formats

- Json
- csv (robust csv) and xls
- txt, docx and pdf

# 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 :

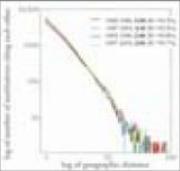
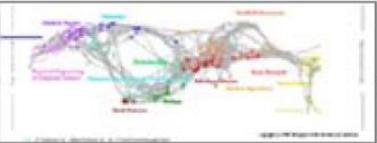
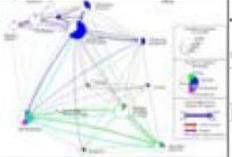
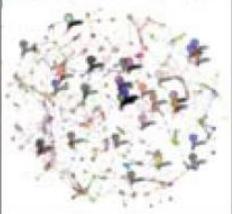
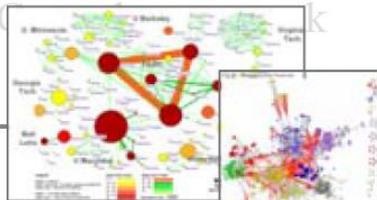
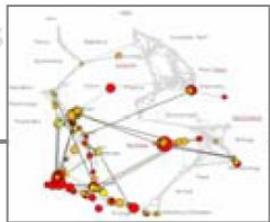
- **Natifs**: qui s'appuient sur des variables directement accessibles dans les données initiales, par exemple dans des métadonnées. Ces indicateurs ne rentrent donc pas dans le contenu des documents.

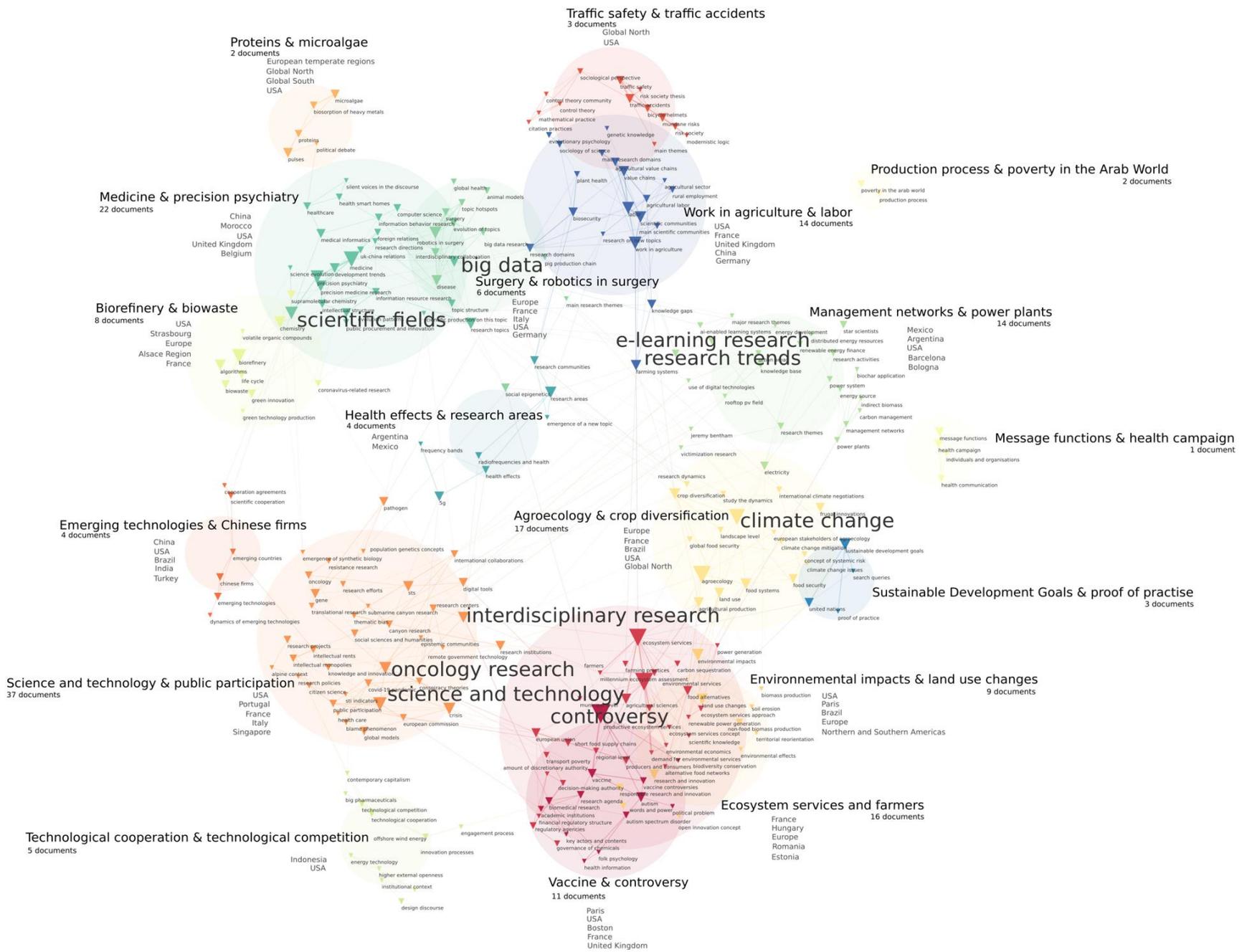
*Par exemple : collaboration entre les auteurs...*

- **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...*

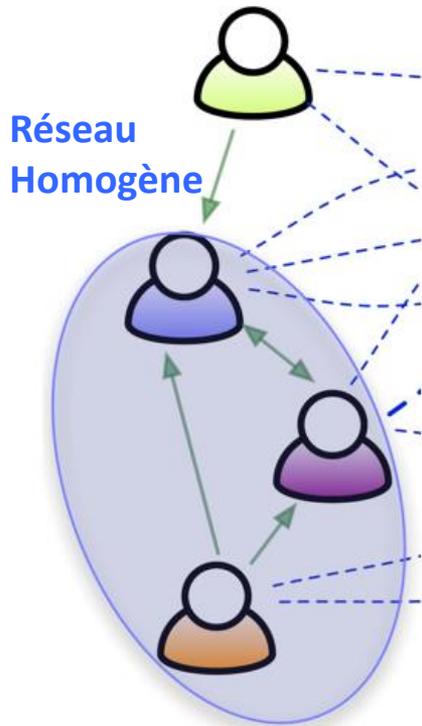
# Dimensions d'analyse et échelles

	<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 

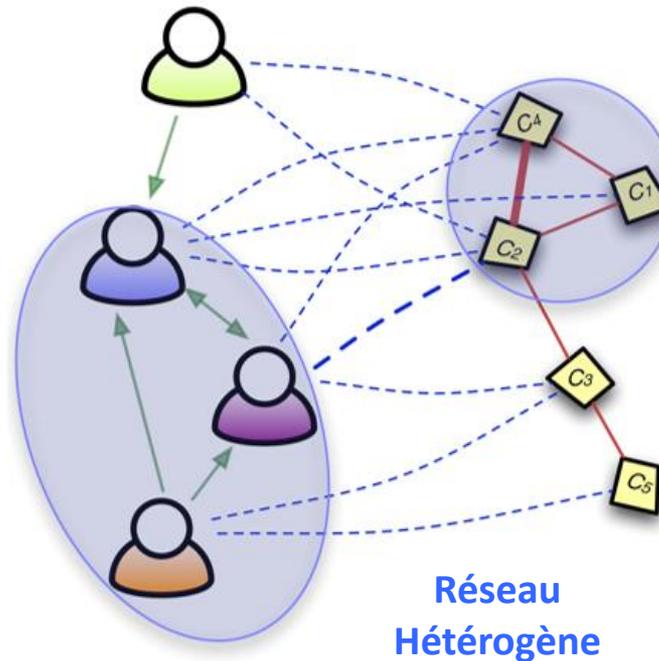


→ <https://www.cortext.net/analysis-of-the-scientific-production-that-mentioned-the-use-of-cortext-manager/>

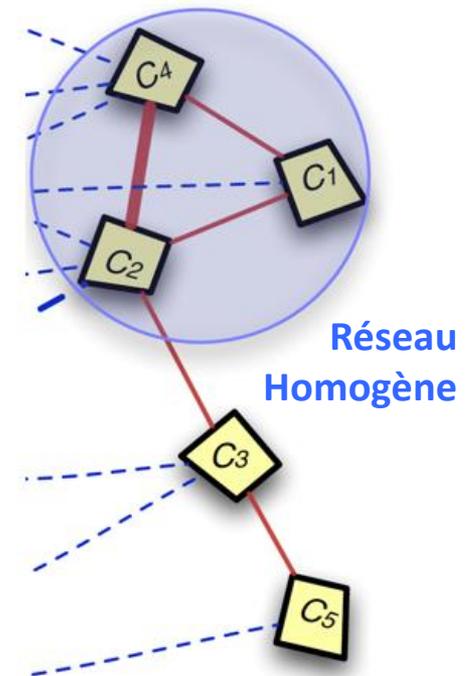
# Combiner relations directes et indirectes : réseau hétérogène



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

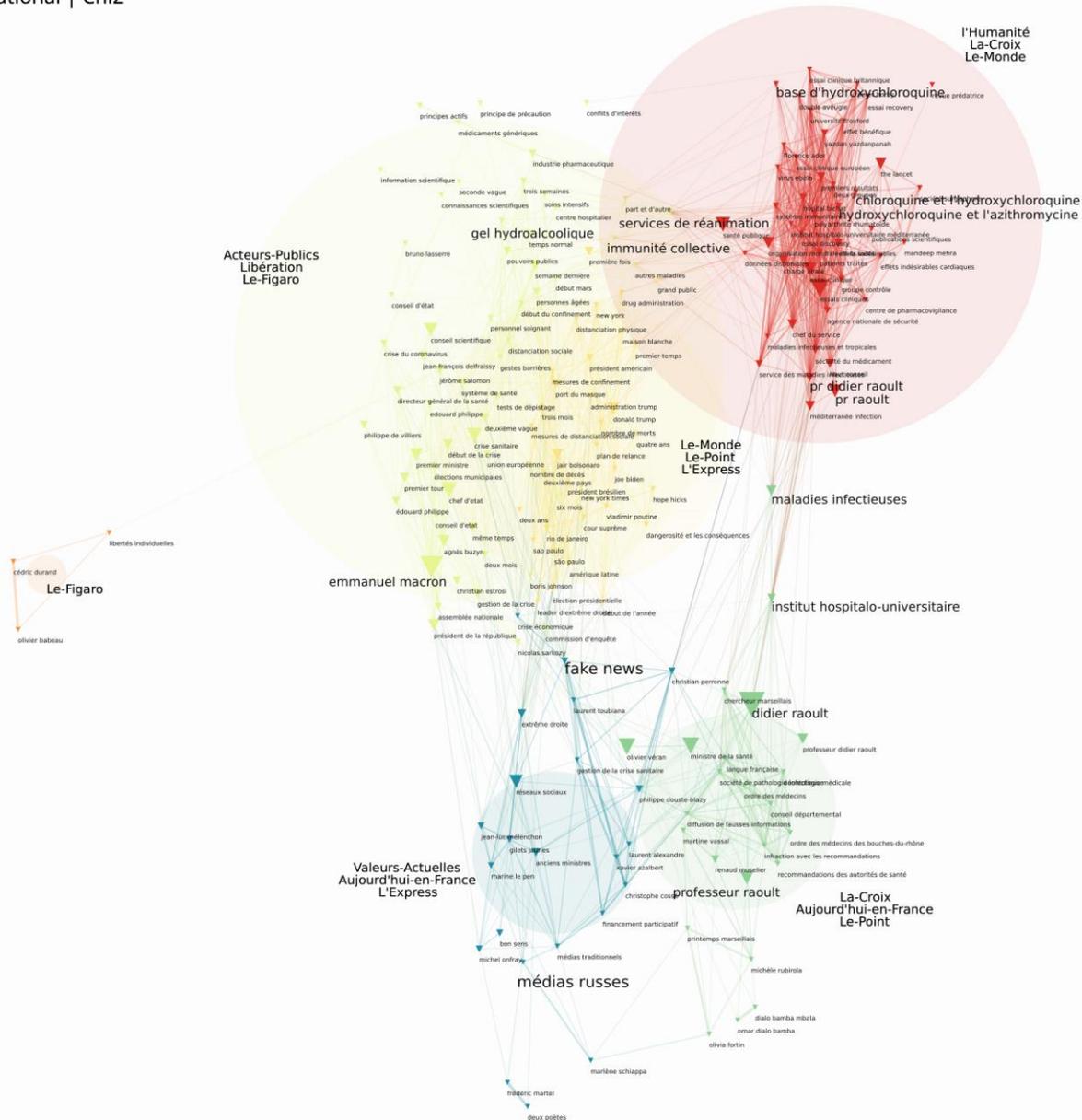
# Les mesures de proximités disponibles

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

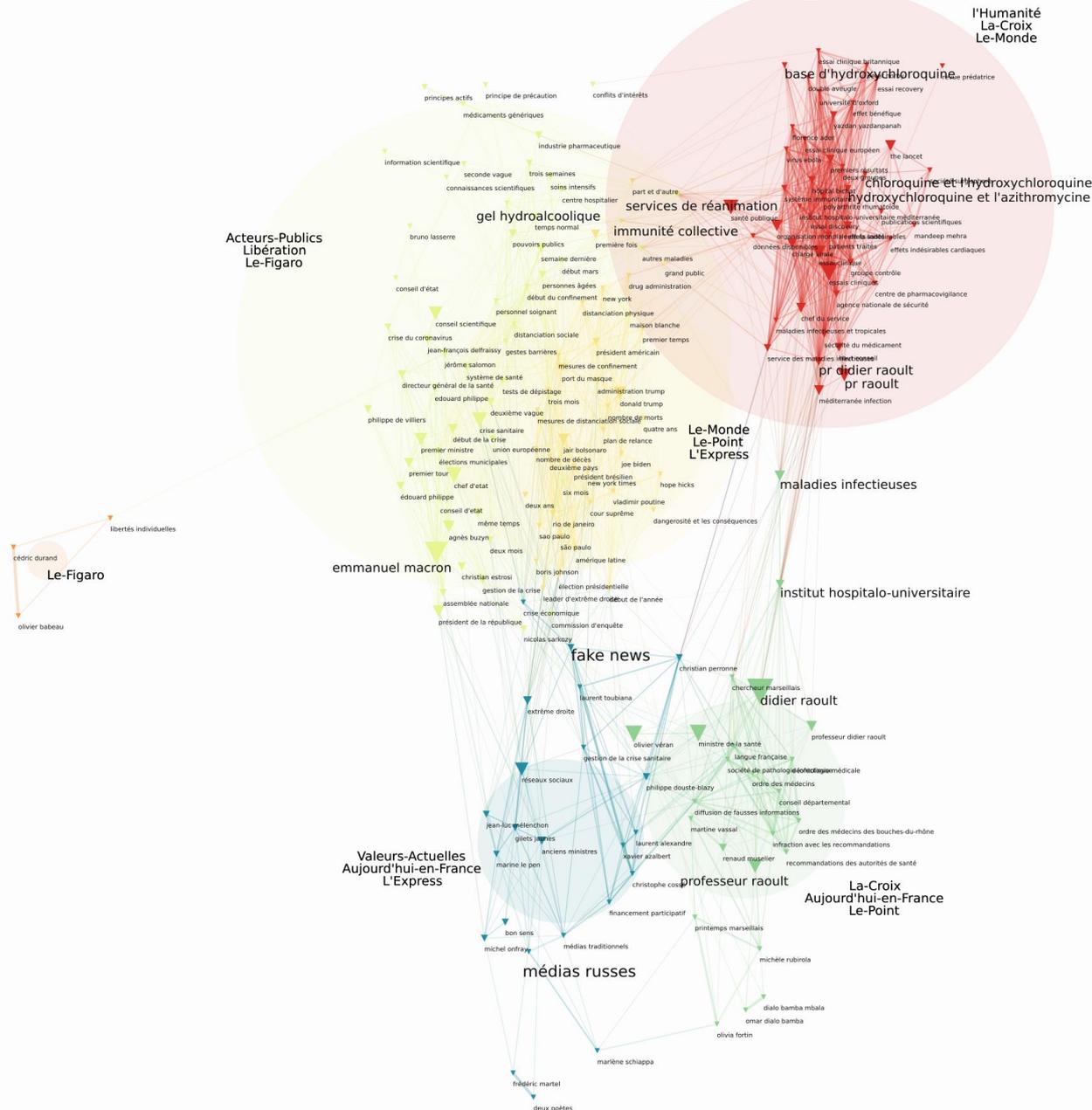


Nombre de clusters  
(espaces sémantiques)

Exemples de métriques

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

# Lecture mésoscopique

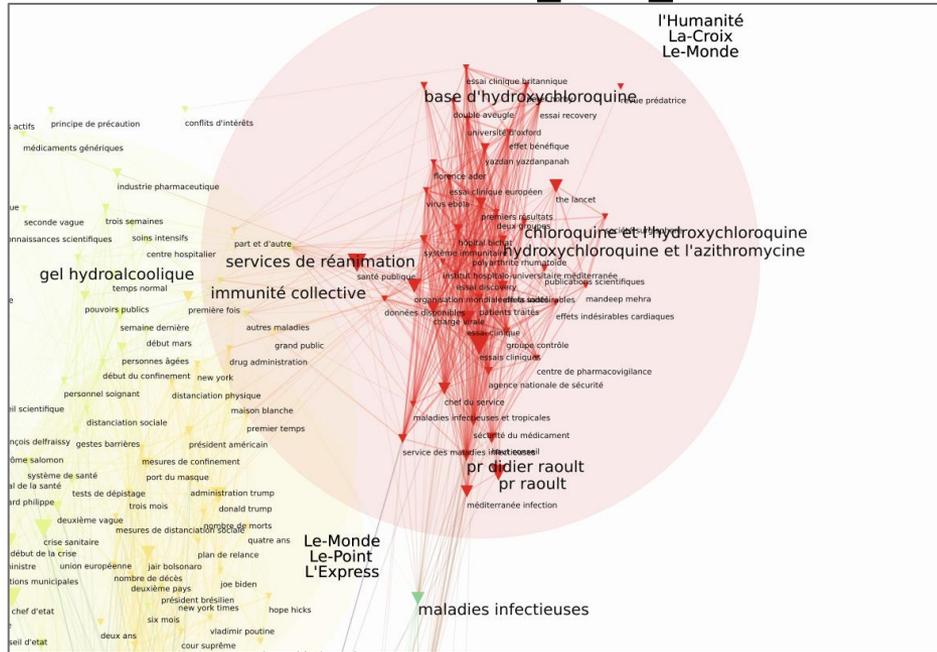


Proximité entre les clusters  
(espaces sémantiques interstitiels)

Exemples de métriques

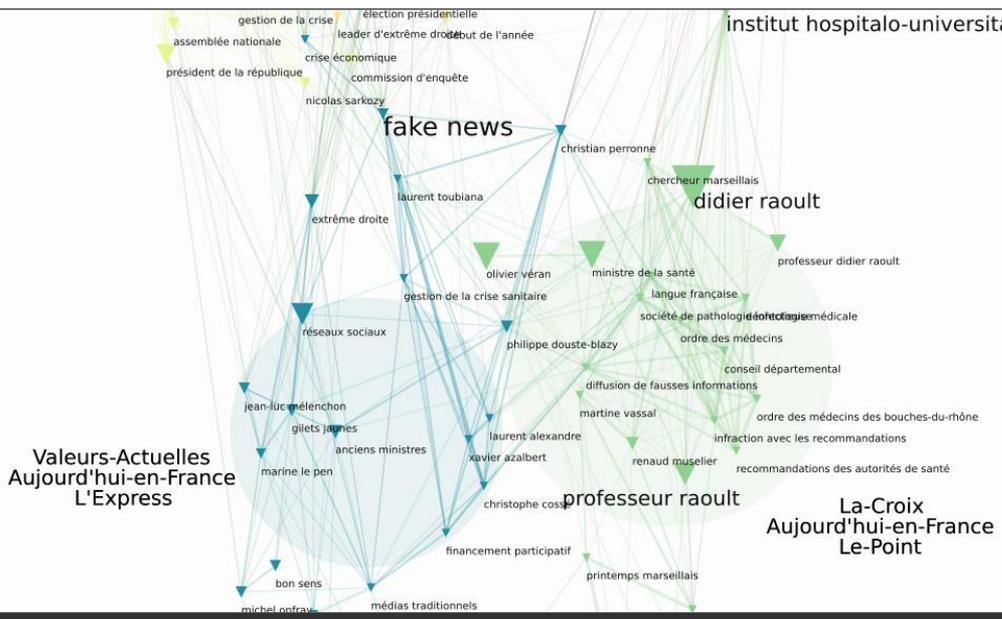
- Distances entre les clusters
- Nombre de liens ou de documents partagés
- Tailles et importances relatives des clusters
- Densités intra-clusters

# Lecture microscopique



Interprétation locale de la composition des clusters

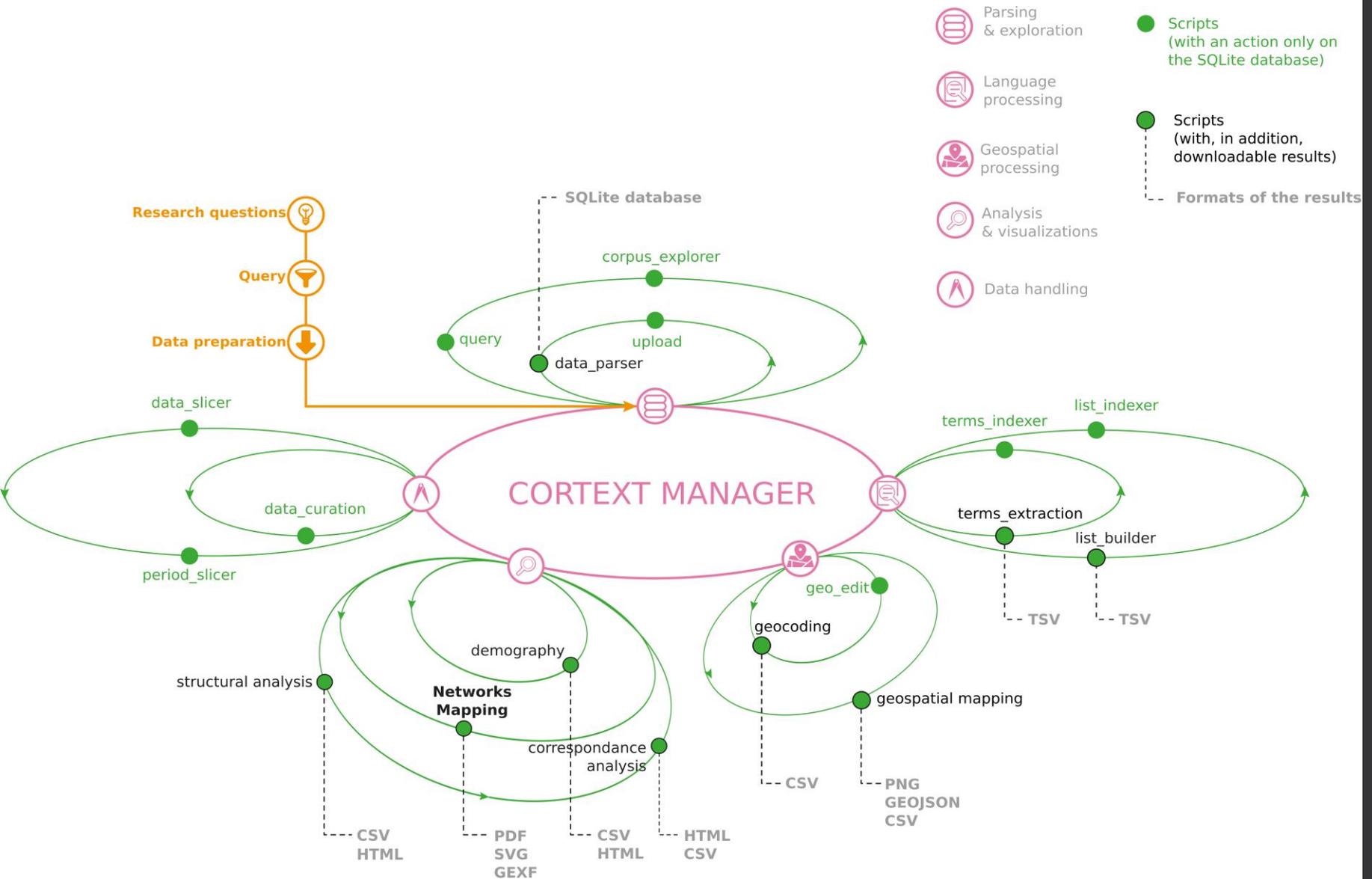
Et des positions des nœuds



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/>

→ <https://docs.cortext.net/trainings/cortext-lisis-2024/>

→ Et

→ <https://managerv2.cortext.net/project/205610003210>