

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

TDM, GTSO données, Consortium Couperin, Juin 2023

Lionel VILLARD

LISIS, IFRIS, INRAe, CorTextT, ESIEE Paris

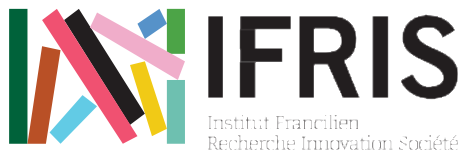
La plateforme CORTEXT

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



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
[✉](#) [t](#) [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.
[t](#) [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
[t](#) [in](#) [ig](#)



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



Hajar Laglil
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
[t](#) [in](#) [ig](#)



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
Bullscience



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**



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

Presse / newspapers

- Factiva
- Europress
- Guardian

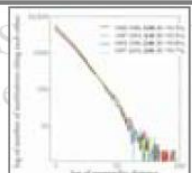
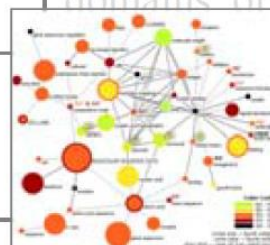



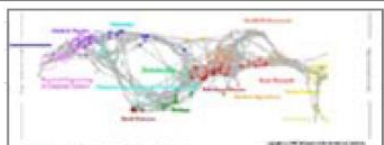

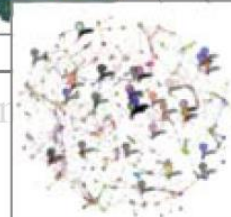

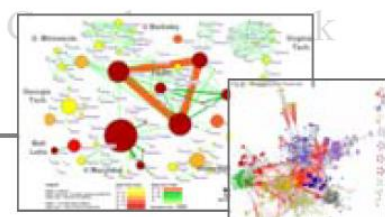
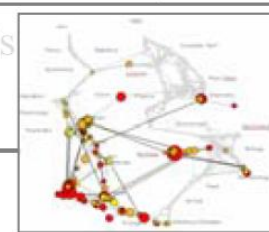
Social media

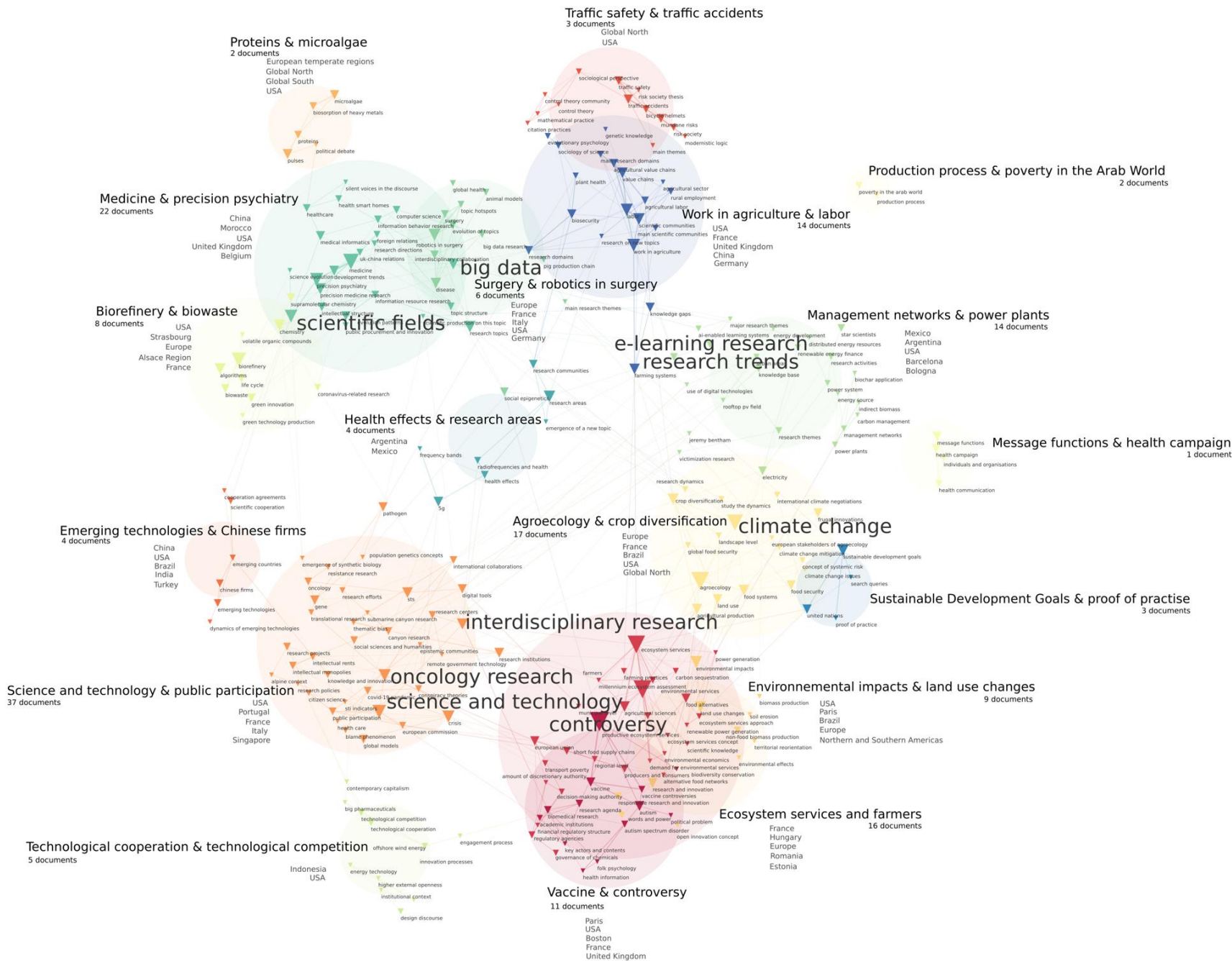
- Twitter : json

Generic formats

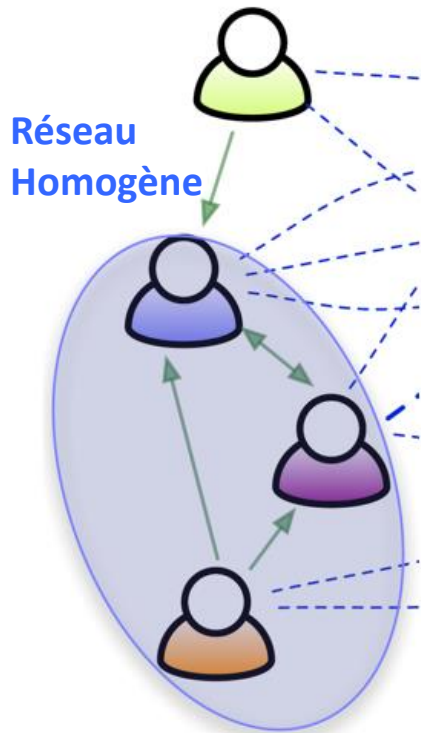
- Json
- csv (robust csv) and xls
- txt, docx and pdf

Dimensions d'analyse et échelles

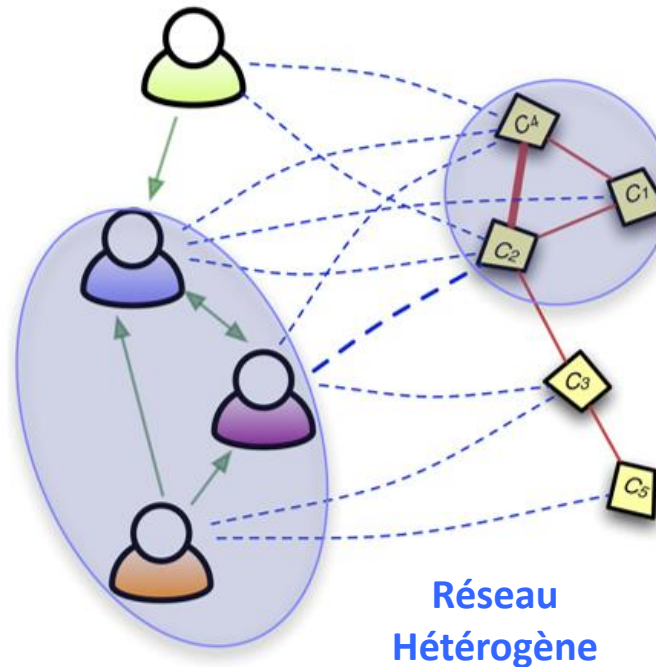
	<i>Micro/Individual (1-100 records)</i>	<i>Meso/Local (101-10,000 records)</i>	<i>Macro/Global (10,000 < records)</i>
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of science in USA, 
Temporal Analysis (When)	Funding portfolio of one individual	Research bursts of PNAS 	113 Years of P Research 
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a scientist's intellectual life 	PNAS 
Topical Analysis (What)			VxOrd/Topic research NIH funding 
Network Analysis (With Whom?)	NSF work of one 	NIH's work of one 	NIH's work of one 



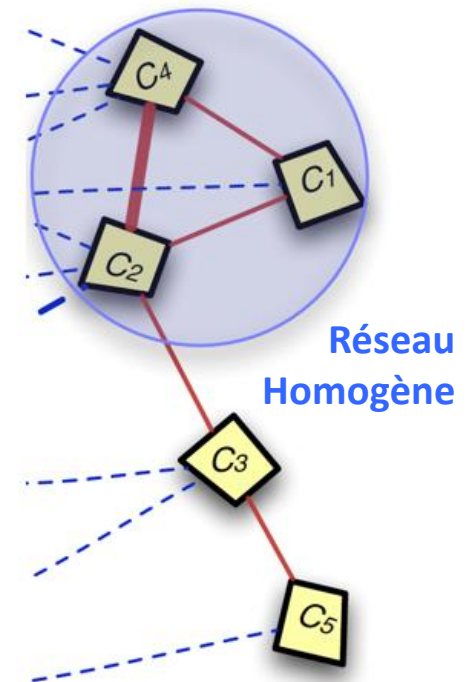
Combiner relations directes et indirectes : réseau hétérogène



Des humains ont des relations:
graphe sociologique



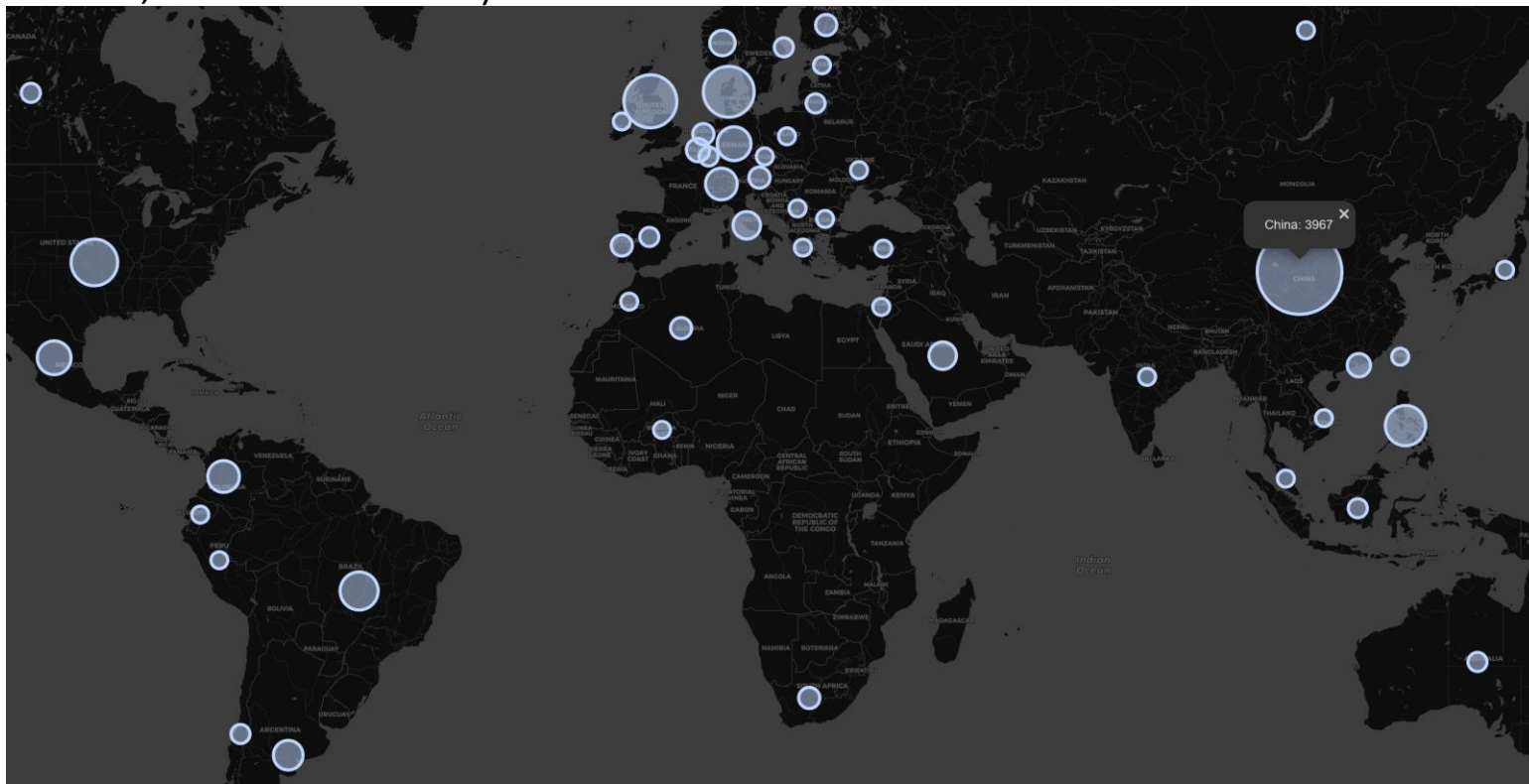
Des humains et des termes ont des relations:
graphe hétérogènes (socio-sémantiques)



Des termes sont associés dans des phrases:
graphe textuel

CorText Manager en 2022

- 1150 utilisateurs actifs, générant plus de 50000 calculs;
- provenant d'environ 456 institutions (universités, entreprises, ministères, cabinets de conseil, journaliste, services de veille...) et villes



Les publications de CorText Manager

- Plus de 890 auteurs ont publié en revendiquant une utilisation de CorText Manager depuis 2016. Ces auteurs représentent environ 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

Avec quatre modes d'usage académique et dans l'enseignement supérieur

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**;
 - et analyse **scientométrique**;
 - ...et des **media sociaux**.
 - Etude du **positionnement relatif** et de la **spécialisation** d'acteurs (e.g. entreprise, université) et d'espaces géographiques (e.g. ville, région, pays)
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
 - **Analyse des réseaux sociaux (SNA)** mettant en lumière des structures organisationnelles et sociales présentes dans les situations observées.
3. **Revue de la littérature** : certains auteurs utilisent CorText Manager pour circonscrire et situer leurs travaux (e.g. parcours doctoral, dans un papier).
4. **Contextes pédagogiques** : traitement automatique des langues et analyse des textes, étude de controverses, étude des dynamiques sociales et politiques sur les média sociaux (e.g. tendances linguistiques, mobilisations sociales, discriminations systématiques), Social Network Analysis, veille technologique et scientifique...

Eclairages stratégiques et rapports d'expertise (avec peu ou pas de traces numériques)

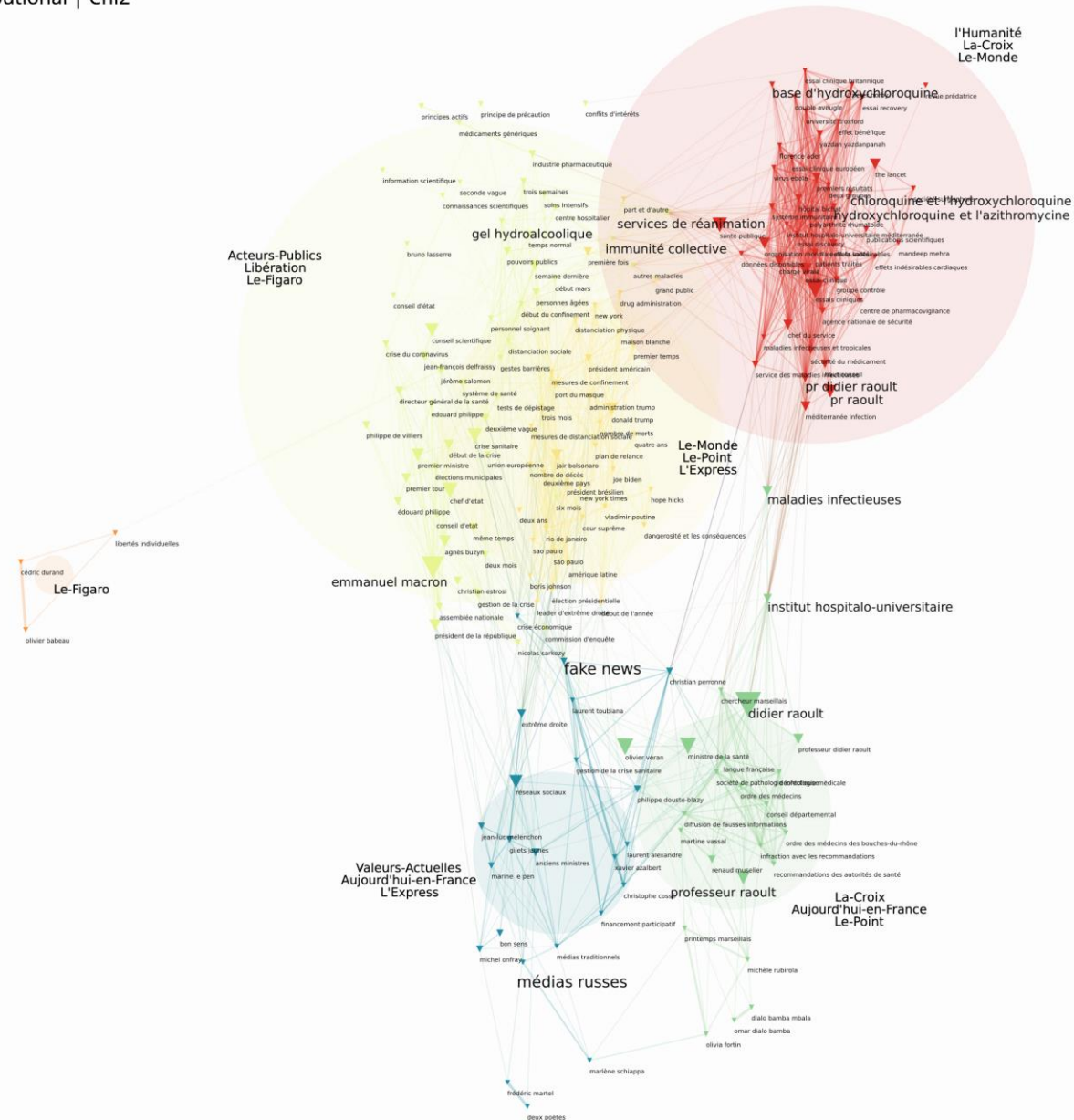
1. **Eclairage politique ou stratégique** : note stratégique à destination du politique, de la conduite de la R&D, politique de la recherche, suivi de "trending topics"...
2. **Rapport d'expertise** : contexte de controverse, fabrique de la norme...

→ <https://docs.cortext.net/trainings/cortext-sciencesconf-2023/04-exemples-papiers/>

Visualisations et les trois niveaux de lecture

Lecture macroscopique

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

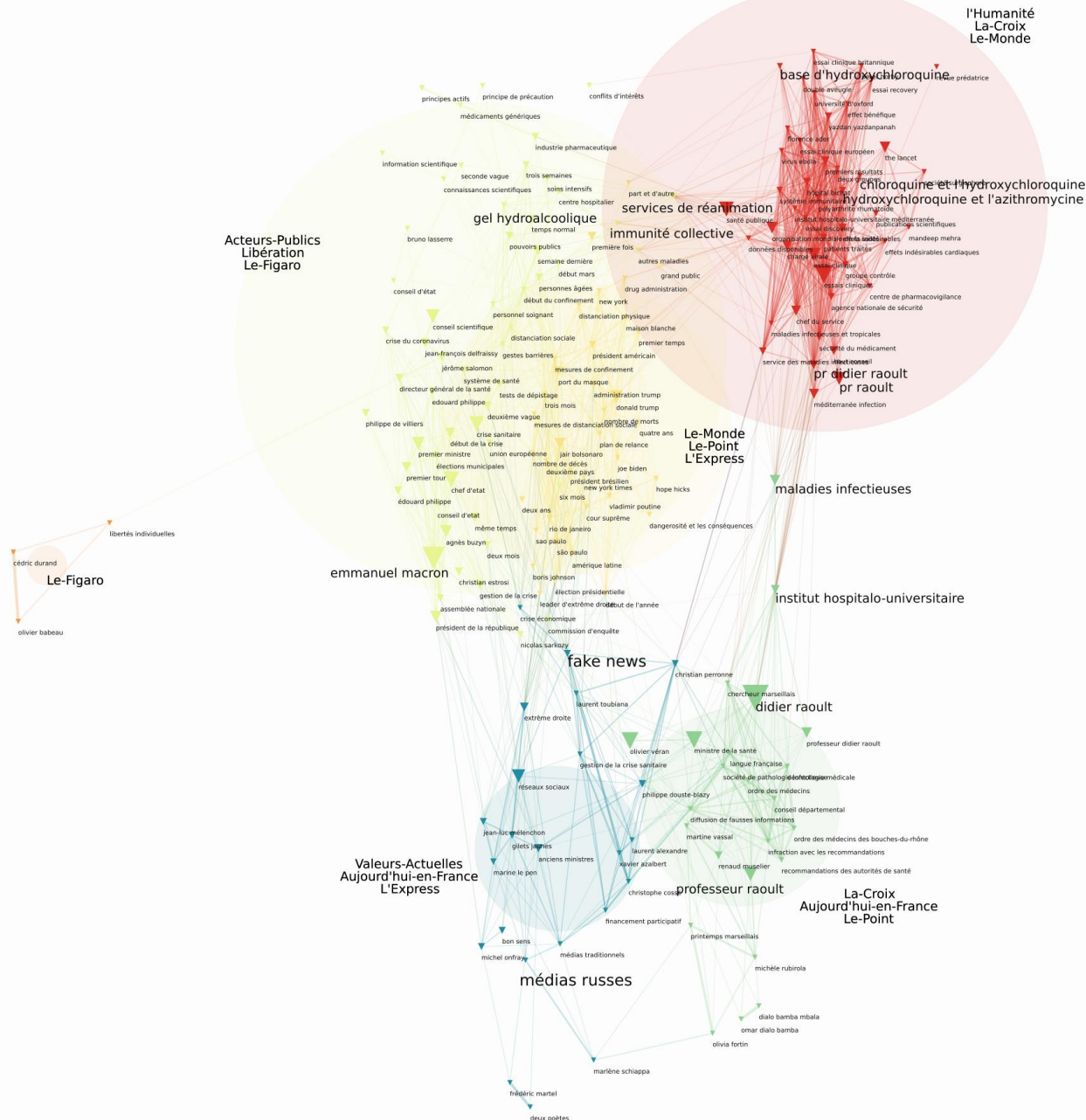


Nombre de clusters
(espaces sémantiques)

Exemples de métriques

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

Lecture mésoscopique

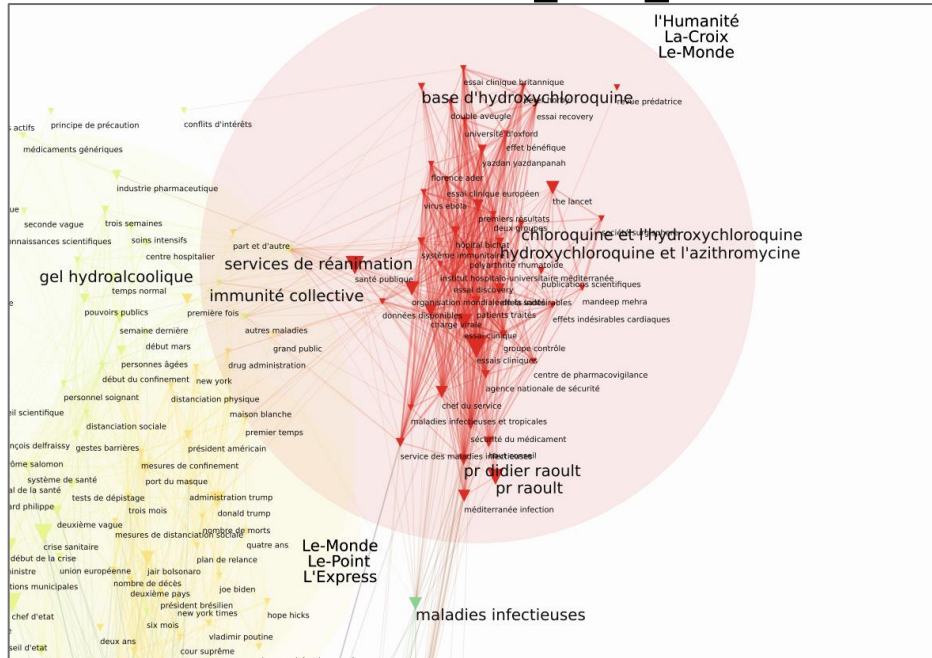


Proximité entre les clusters (espaces interstitiels) sémantiques

Exemples de métriques

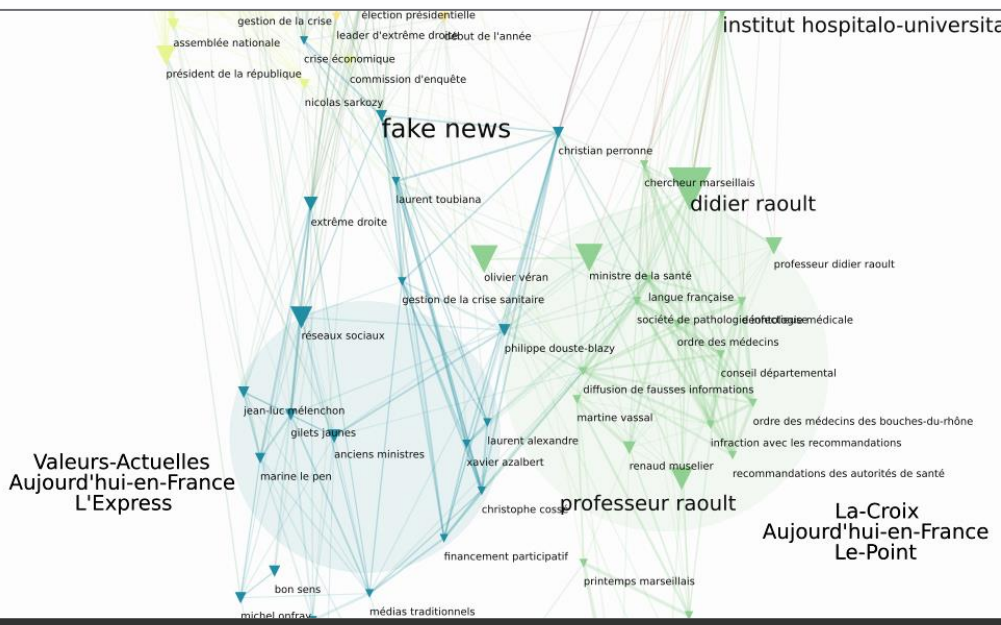
- Distances entre les clusters
- Nombre de liens ou de documents partagés
- Tailles et importances relatives des clusters
- Densités intra-clusters

Lecture microscopique



Interprétation locale de la composition des clusters

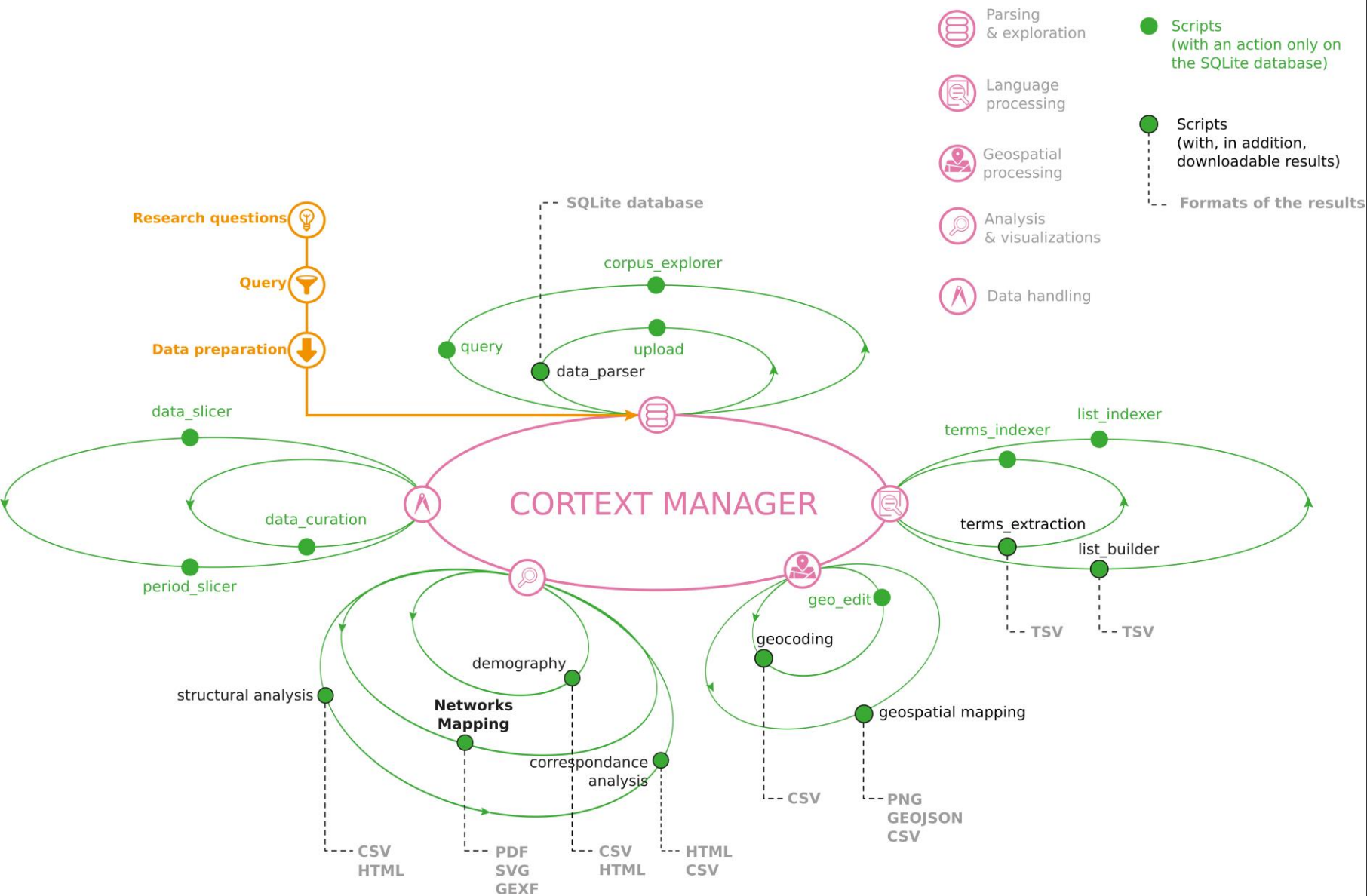
Et des positions des nœuds



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-sciencesconf-2023/>

Et

→ <https://managerv2.cortext.net/project/169270003210>