

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

Tuto@Mate, Décembre 2022

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.**

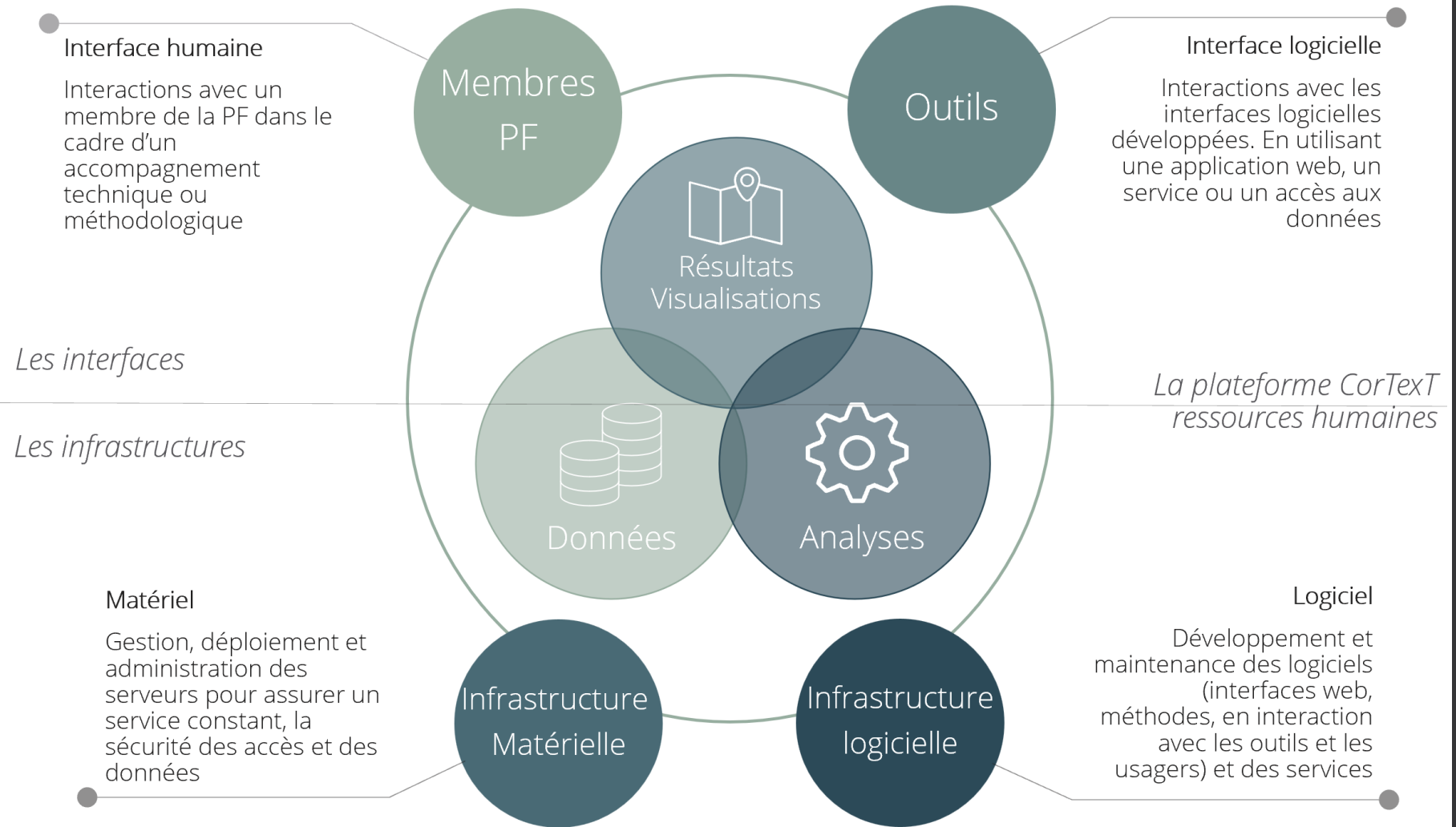
*“L’objectif premier de la plateforme CorText est qu’un chercheur en sciences sociales – ou un autre utilisateur – puisse venir avec une **question de recherche** et partir en **profitant des fruits d’une méthode computationnelle** adaptée à sa question.”*



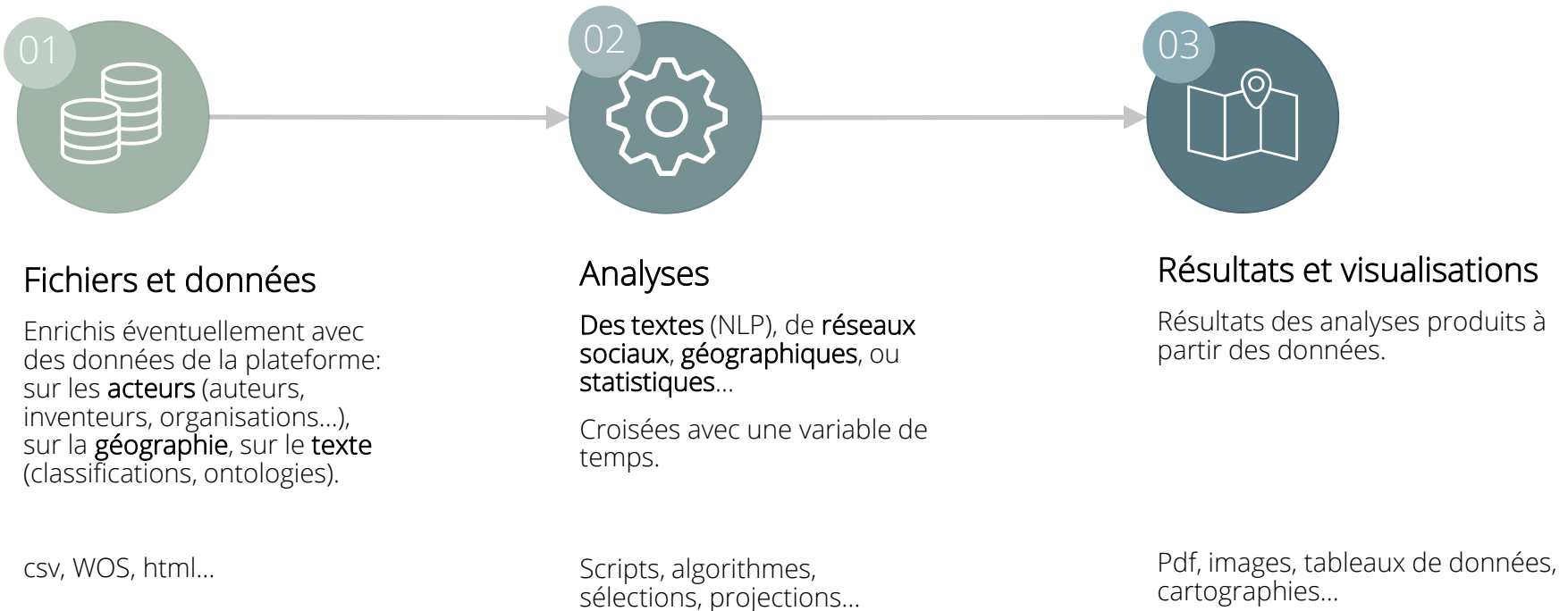
avec les soutiens de

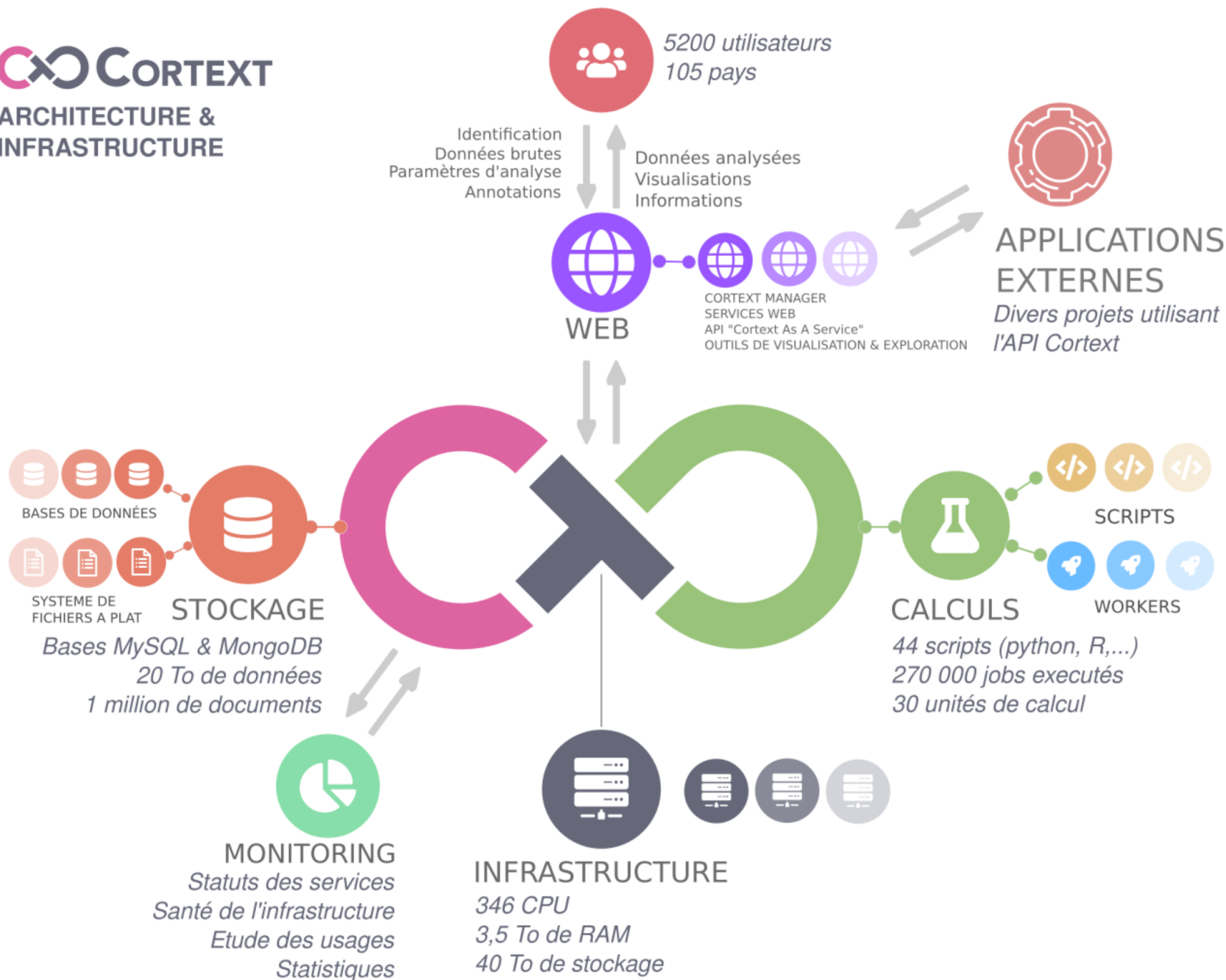


La plateforme CorTextT : entre interfaces et infrastructures



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





**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

**Pierre-Yves Bulot**

IT Engineer Assistant, CorText

**Luis-Daniel Medina**

IT Engineer, CorText

**Diego-Fernando Gómez Peña**

IT Engineer, CorText

**Tatiana Andrea Sánchez Castaño**

IT Engineer, CorText

**Joenio Marques da Costa**

Research Software Engineer, CorText

**Géraldine Enderli**Engineer specialised in the
production, processing and analysis
of data and survey at INRAE - LISIS,
CorText**Hajar Laglil**Meteorological engineer/Data
scientist, CorTextINTERACTION CHERCHEURS,
PROJETS, FORMATIONS,
VALORISATION, COMMUNICATION5 CHERCHEURS
1 ASSISTANT INGÉNIEURCOLLECTE 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, SCRIPTS1 INGÉNIEUR DE RECHERCHE
2 INGÉNIEURS D'ETUDESGESTION 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**Fondatrice 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

**Loïc Boudoulec**

IT Engineer

**Bertha Brenes**

IT Engineer, Trainee, CorText

**Anis Arabi**

Big data engineer

**Nicolas Ricci**

Web developer and data

**Audrey Baneyx**Project manager Data science,
Sciences Po Paris - medialab**Andrei Mogoutov**

Bullscience

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

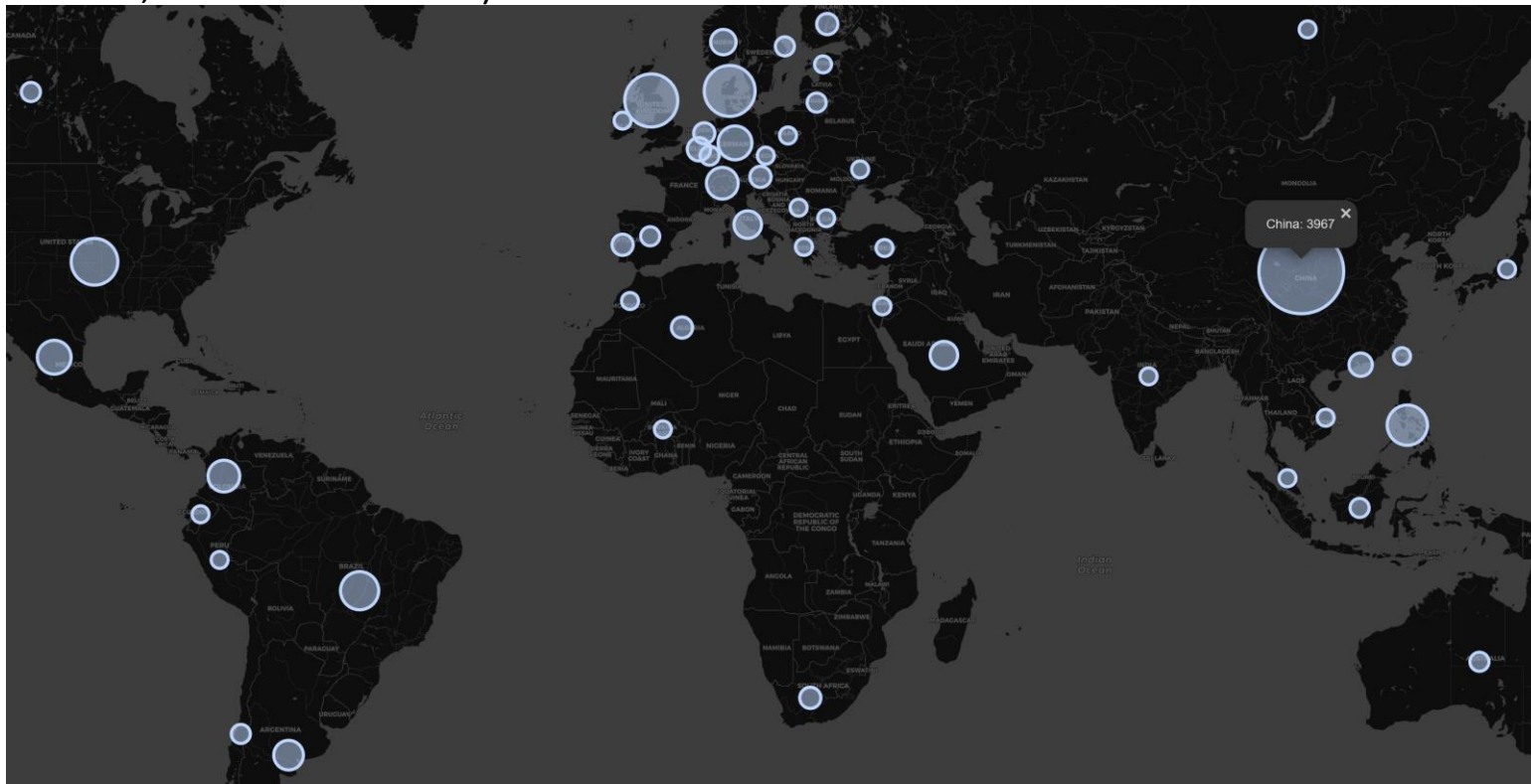
Champs disciplinaires mobilisés

- **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



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

JOB NAME

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

job label (optional)

SCRIPT PARAMETERS

Source

Type of Data

☒ dataset ☐ cortext db

Corpus Format

Ignore entries with

☒ yes ☐ no

isi
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

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



Research Policy 36 (2007) 893–903



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, 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”.

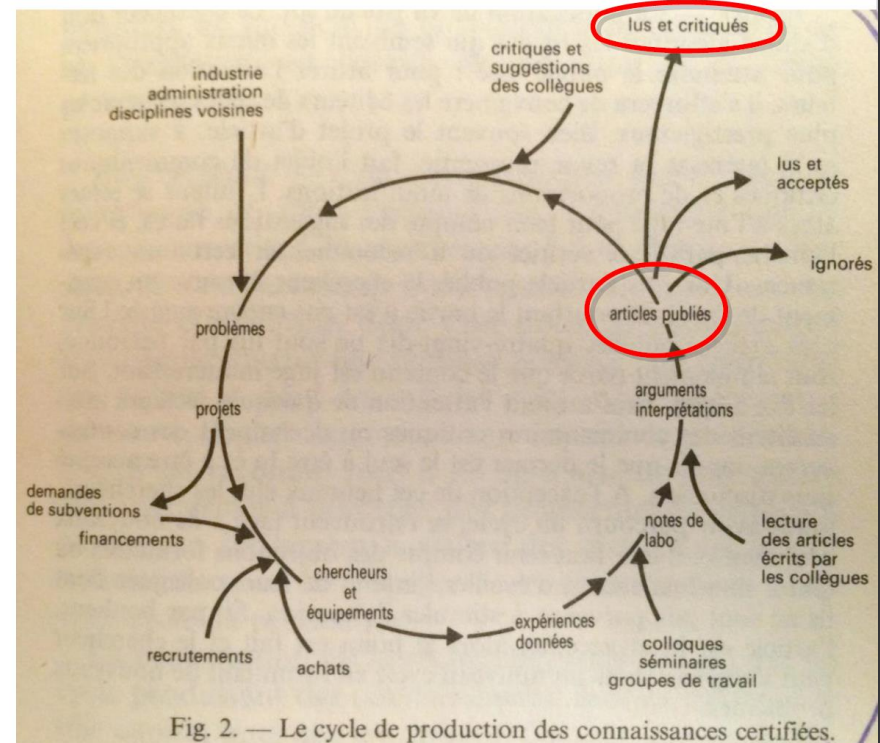
Citations et références de l'article : sources scientifiques de l'article

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 < 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
Temporal Analysis (When)	Funding portfolio of one individual	Scientific bursts of PNAS	113 Years of PNAS Research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a scientific intellectual landscape	PNAS global distribution
Topical Analysis (What)	Research topics of one individual	Research topics of a lab or center	VxOrd/Topic modeling of NIH funding
Network Analysis (With Whom?)	NSF network of one individual	NSF network of a lab or center	NIH's network of a lab or center

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

Modes d'usage identifiés

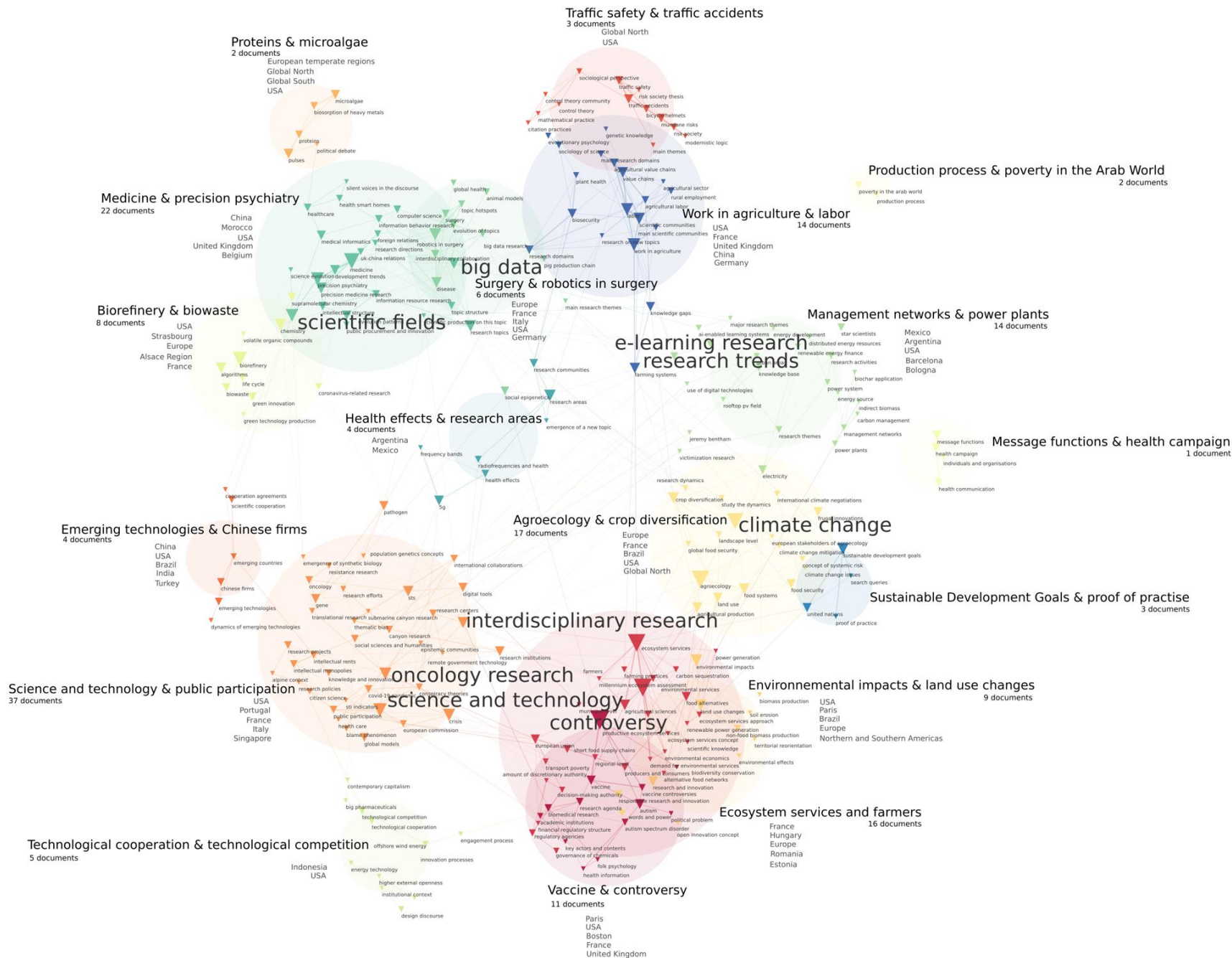
Modes d'usage avec des traces numériques

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).

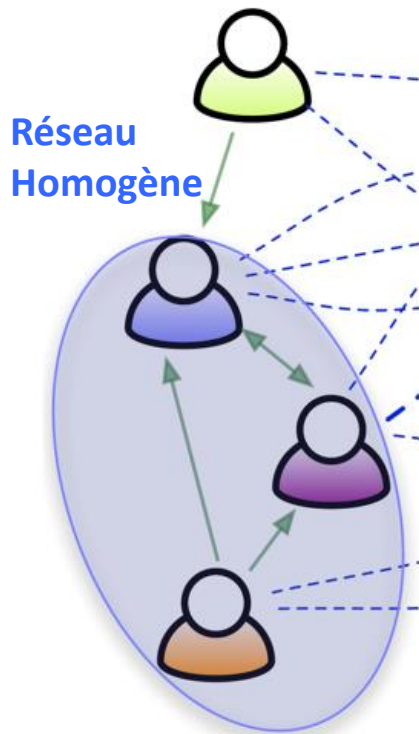
Modes d'usage sans traces numériques

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...
3. **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 systémiques), Social Network Analysis, veille technologique et scientifique...

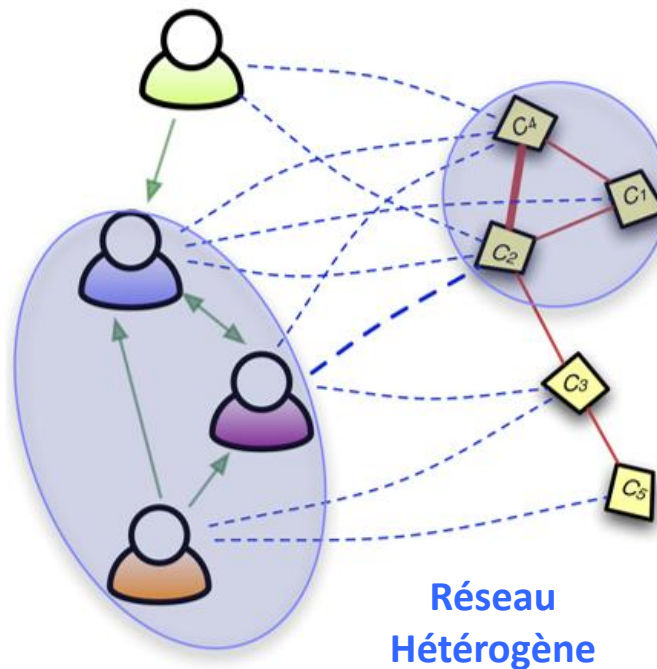
→ <https://docs.cortext.net/trainings/cortext-tutomat-2022/04-exemples-papiers/>



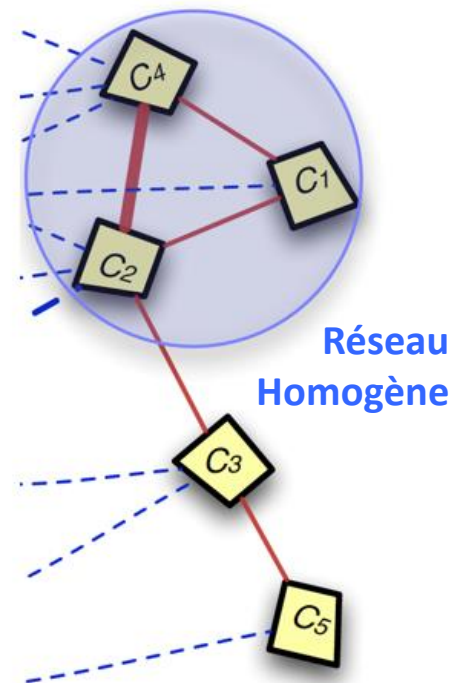
Combiner relations directes et indirectes : réseau hétérogène



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

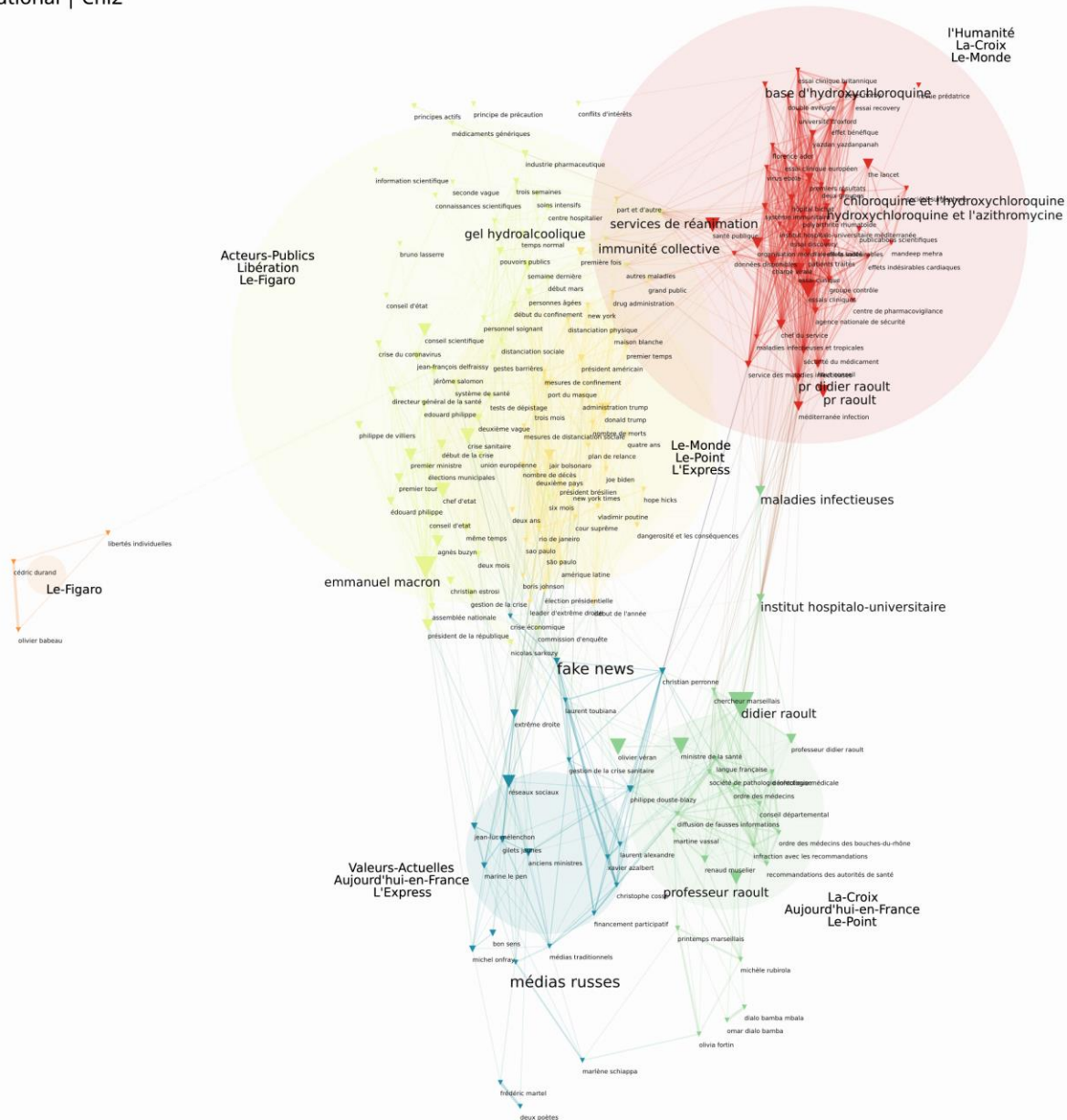
Les mesures de proximités disponibles

proximity measures	type of network	normalisation	special properties
raw	interaction network (e.g. social network)	no	-
χ^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

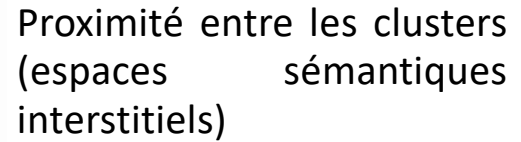


Nombre de clusters
(espaces sémantiques)

Exemples de métriques

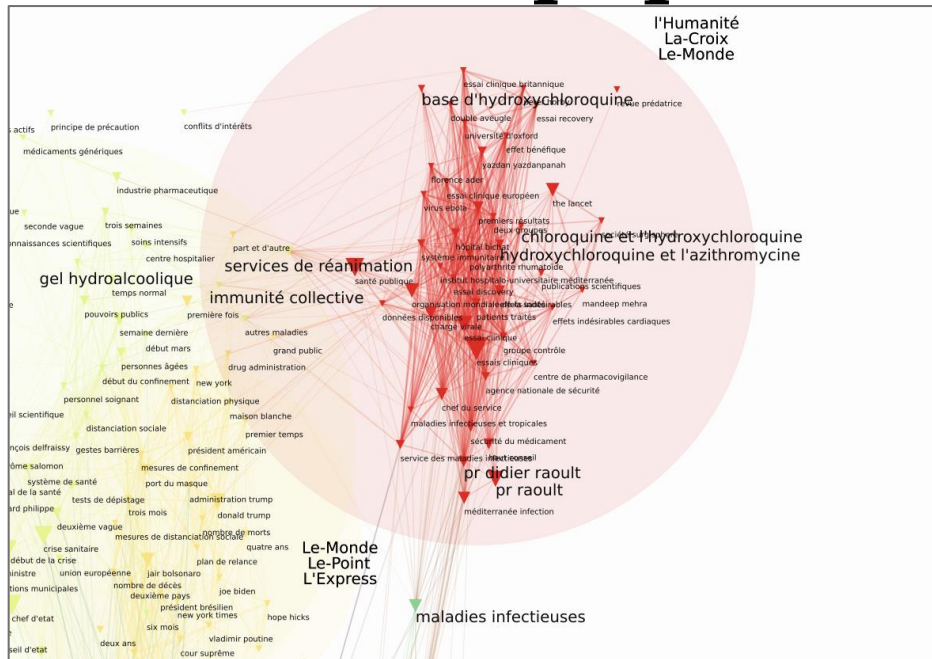
- Nombre de clusters
- Densité du réseau
- Réseau centralisé ou distribué
- ...

Tuto@Mate | Introduction à CorText Manager



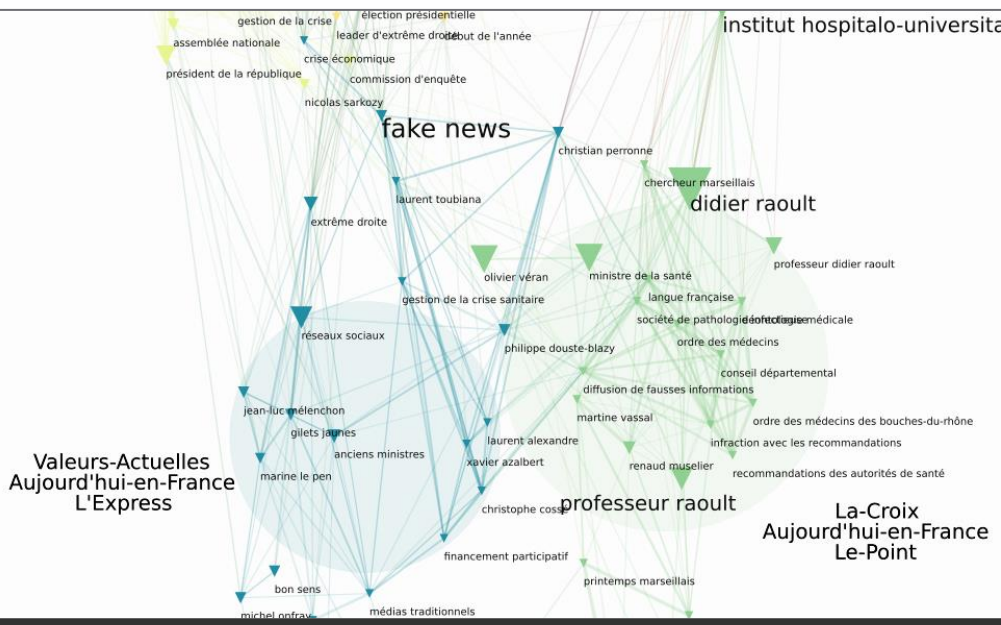
- Distances entre les clusters
- Nombre de liens ou de documents partagés
- Tailles et importances relatives des clusters
- Densités intra-clusters

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

Tuto@Mate | Introduction à CorText Manager



Aller plus loin



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

Et

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