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

Tuto@Mate, Décembre 2022

Lionel VILLARD

LISIS, IFRIS, INRAe, CorTexT, 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."*

LISIS INRAE ESIEE PARIS Université Gustave Eiffel

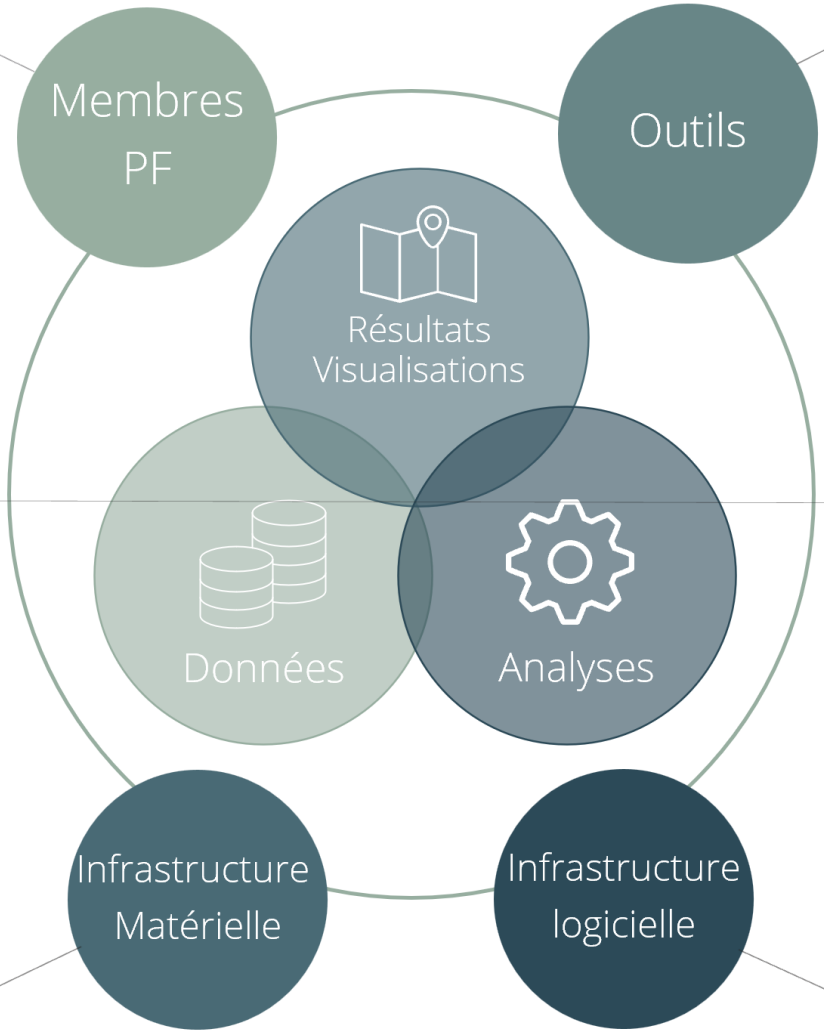
avec les soutiens de

IFRIS Institut Francilien Recherche Innovation Société

LabEx Science, Innovation and TEchnology in Society

RISIS RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

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

Interface logicielle

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

Les interfaces

La plateforme CorTexT ressources humaines

Les infrastructures

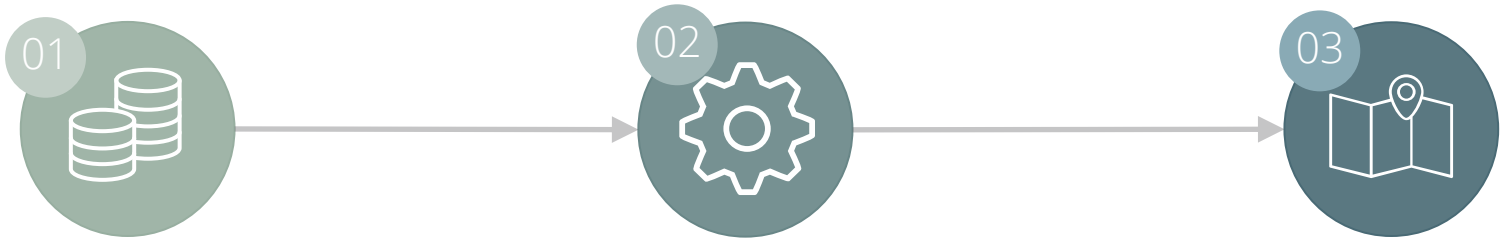
Matériel

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

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)



Fichiers et données

Enrichis éventuellement avec des données de la plateforme: sur les **acteurs** (auteurs, inventeurs, organisations…), sur la **géographie**, sur le **texte** (classifications, ontologies).

csv, WOS, html…

Analyses

Des textes (NLP), de **réseaux sociaux**, **géographiques**, ou **statistiques**…

Croisées avec une variable de temps.

Scripts, algorithmes, sélections, projections…

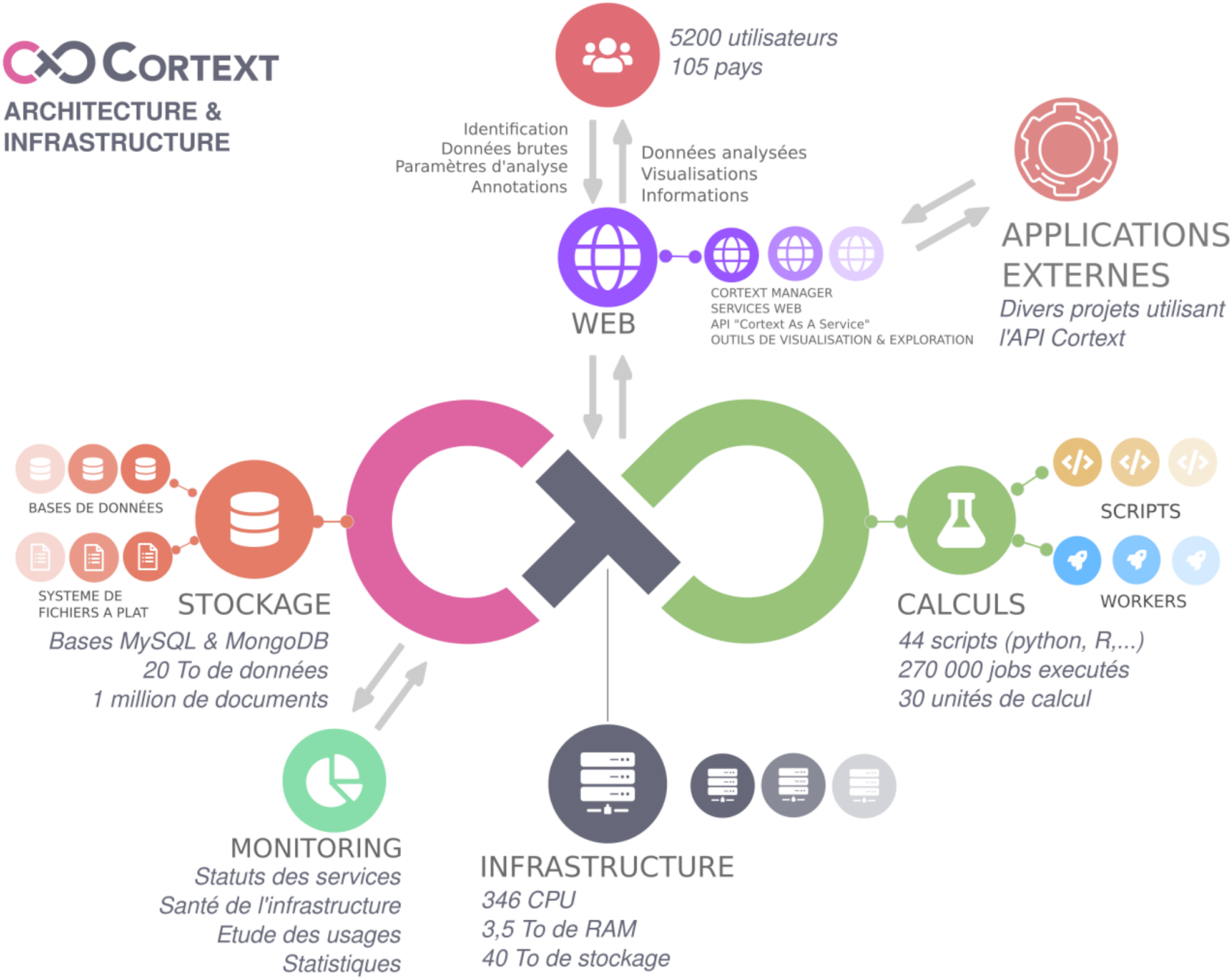
Résultats et visualisations

Résultats des analyses produits à partir des données.

Pdf, images, tableaux de données, cartographies…

CORTEXT

ARCHITECTURE & INFRASTRUCTURE



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, 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

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

Champs disciplinaires mobilisés

- **Scientométrie**
- **Analyse des réseaux sociaux**
- Traitement automatique de la **langue**
- **Statistiques**

- **Cartographie** géographique et **analyse spatiale**

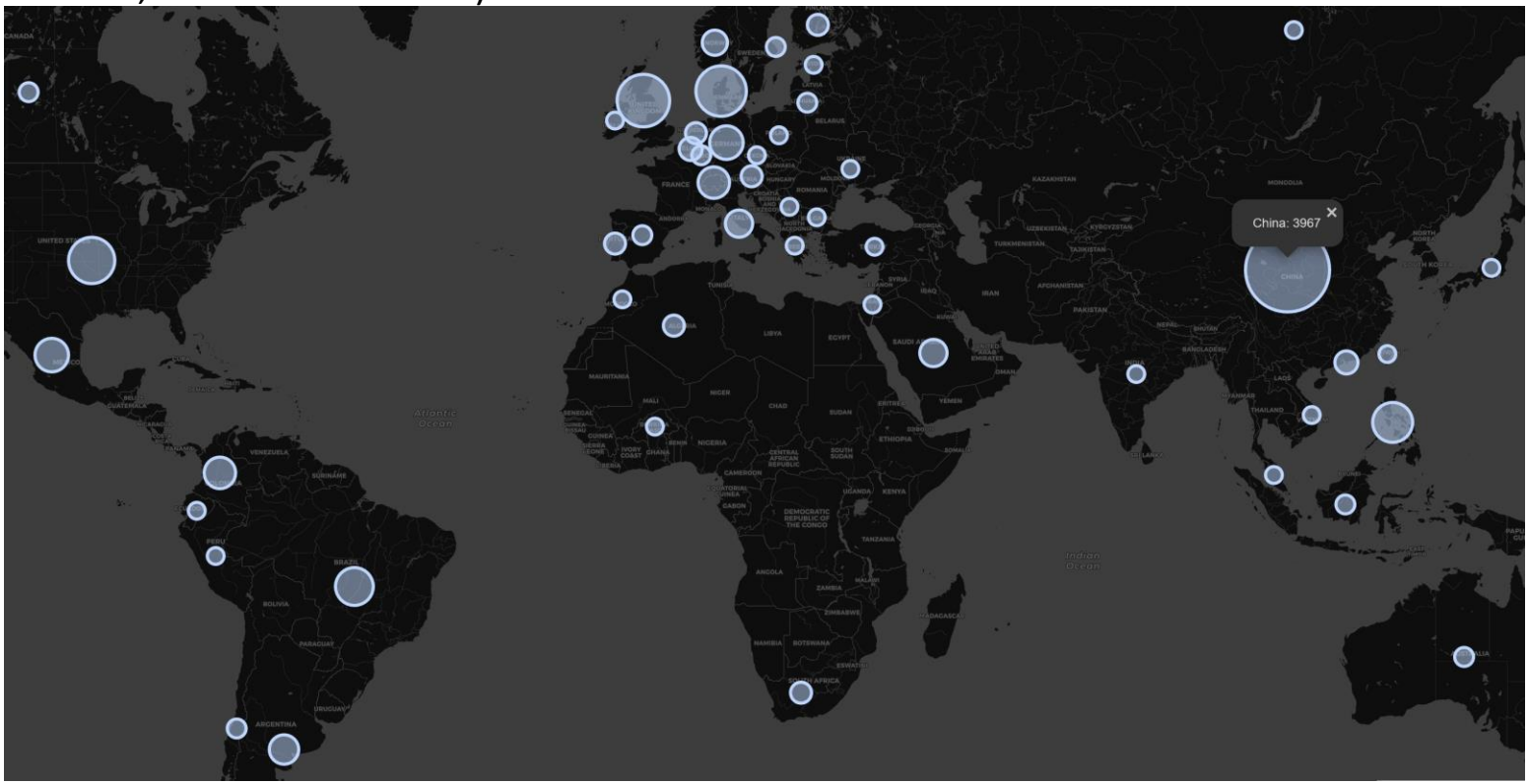
- **Visualisation** de données

- 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/

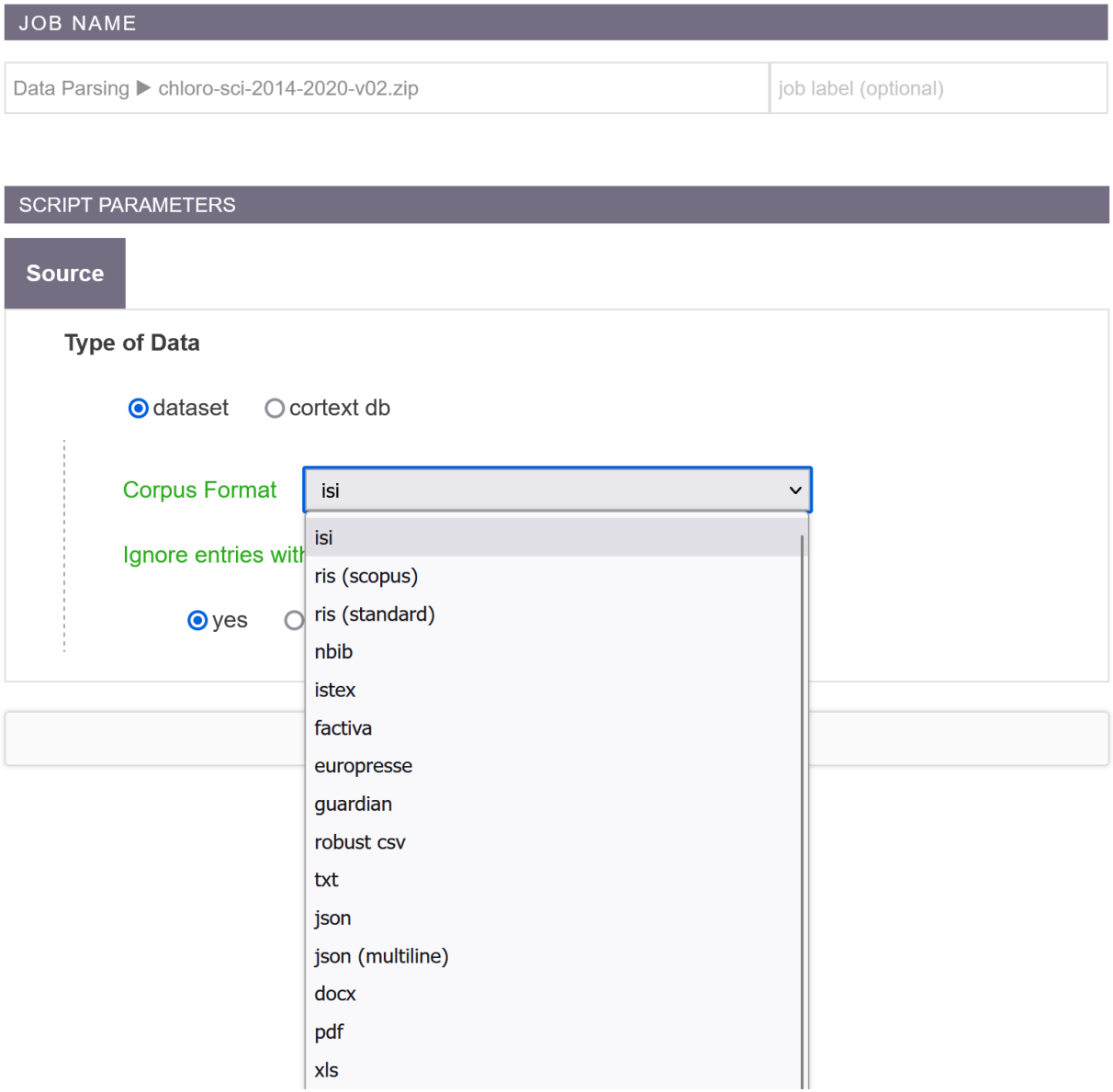
Des jeux de données multiples

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

Presse / newpapers
- 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

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

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 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).

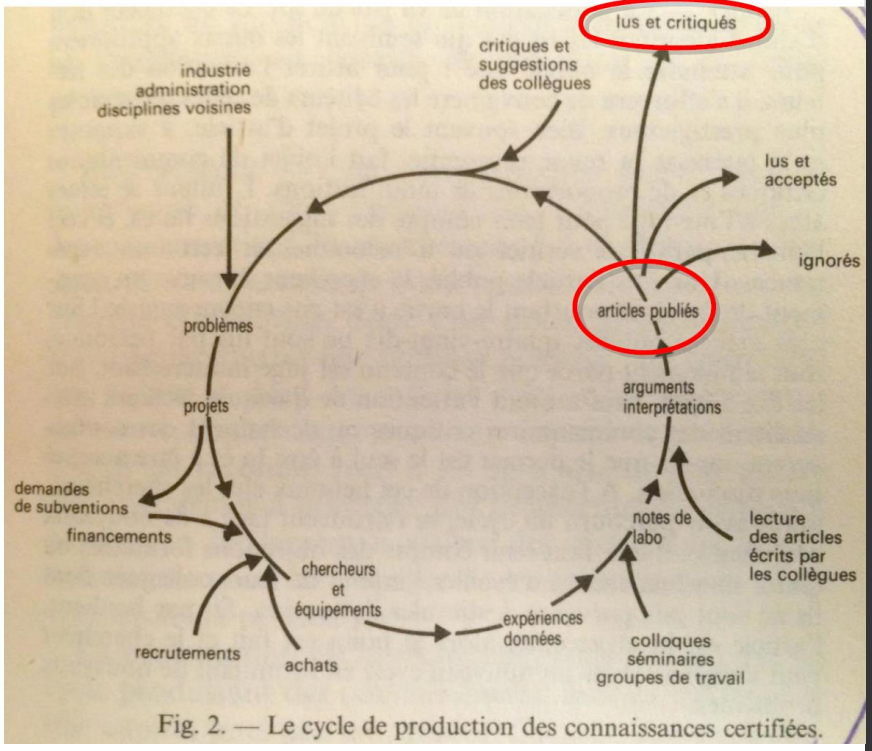
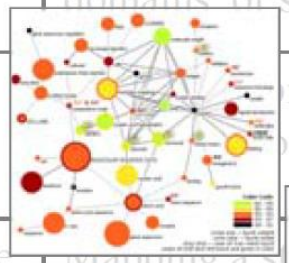
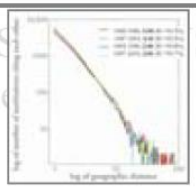


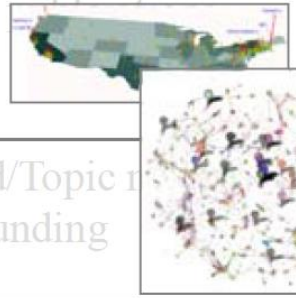
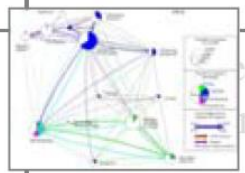
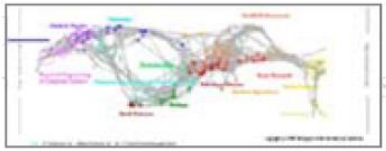
Fig. 2. — Le cycle de production des connaissances certifiées.

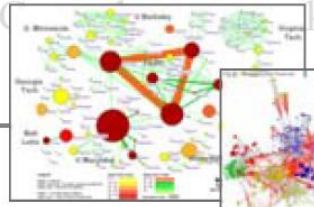
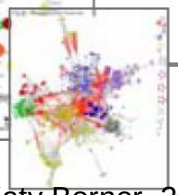
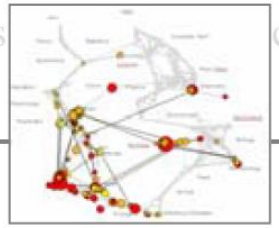
Dimensions d'analyse et échelles

	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.	flows in research	VxOrd/Topic m 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

Modes d'usage identifiés

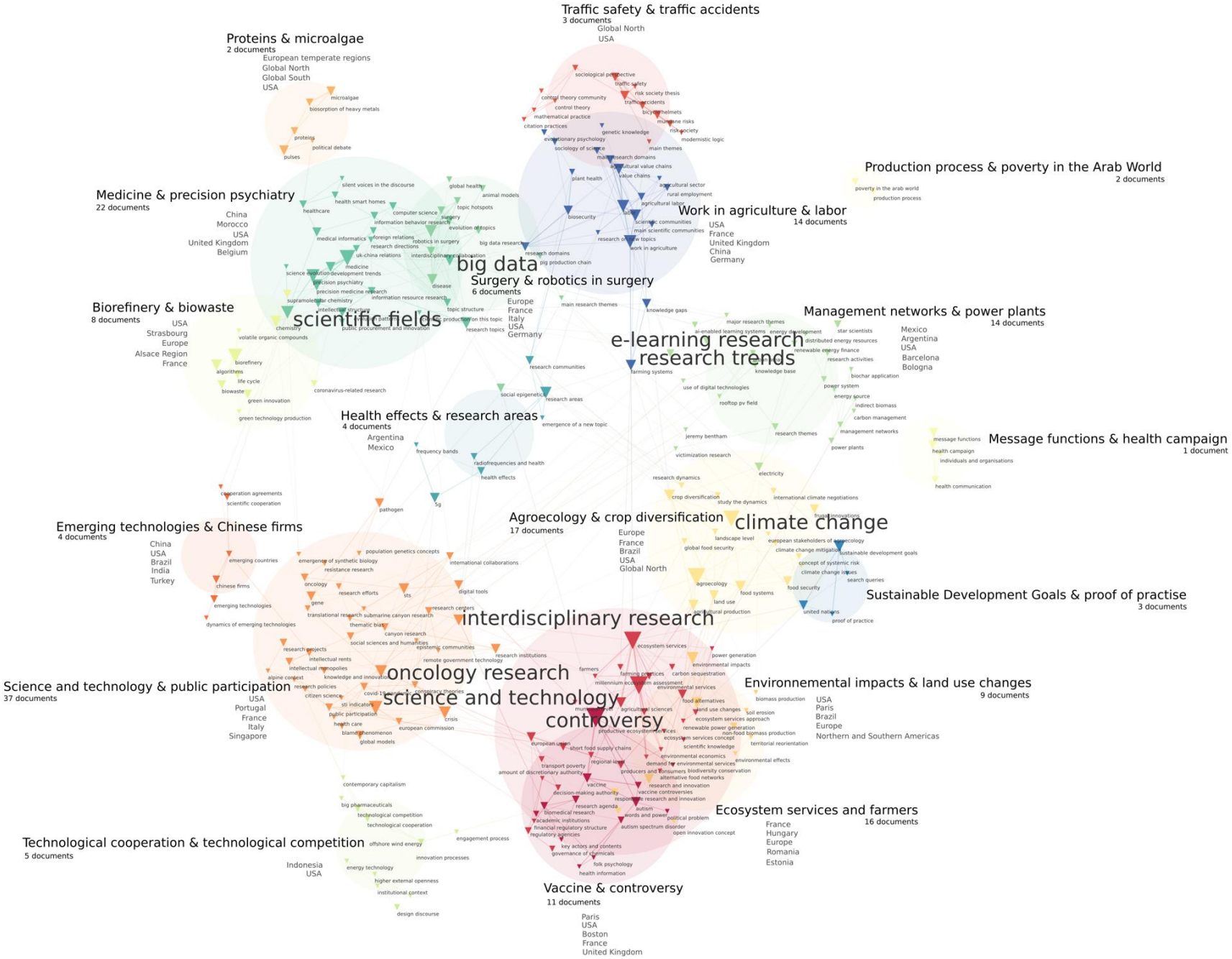
Modes d'usage avec des traces numériques

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).

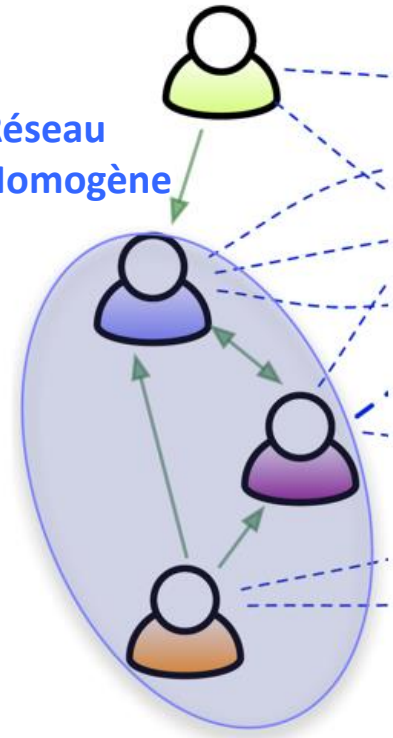
Modes d'usage sans traces numériques

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 systématiques), Social Network Analysis, veille technologique et scientifique...

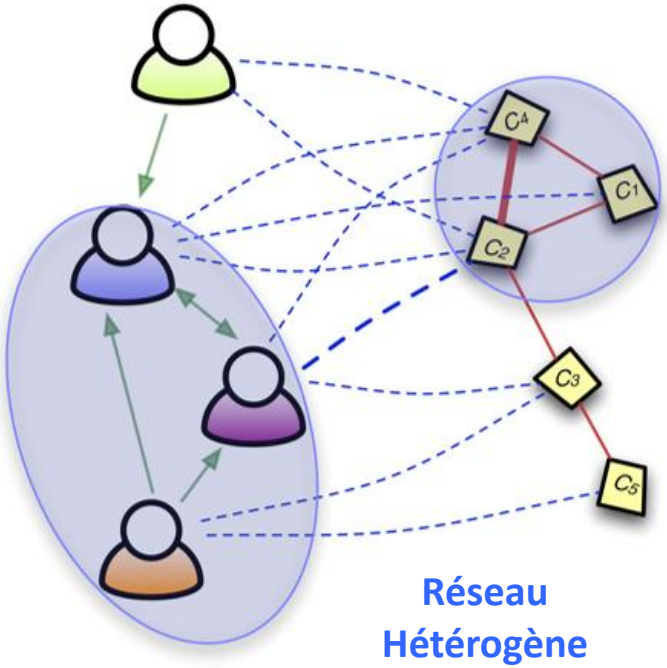
→ https://docs.cortext.net/trainings/cortext-tutomat-2022/04-exemples-papiers/


Traffic safety & traffic accidents
3 documents
Global North
USA
Proteins & microalgae
2 documents
European temperate regions
Global North
Global South
USA
Production process & poverty in the Arab World
2 documents
Medicine & precision psychiatry
22 documents
China
Morocco
USA
United Kingdom
Belgium
Work in agriculture & labor
14 documents
USA
France
United Kingdom
China
Germany
big data
Surgery & robotics in surgery
6 documents
Europe
France
Italy
USA
Germany
Biorefinery & biowaste
8 documents
USA
Strasbourg
Europe
Alsace Region
France
scientific fields
Management networks & power plants
14 documents
Mexico
Argentina
USA
Barcelona
Bologna
e-learning research
research trends
Health effects & research areas
4 documents
Argentina
Mexico
Message functions & health campaign
1 document
Emerging technologies & Chinese firms
4 documents
China
USA
Brazil
India
Turkey
Agroecology & crop diversification
17 documents
Europe
France
Brazil
USA
Global North
climate change
Sustainable Development Goals & proof of practise
3 documents
interdisciplinary research
oncology research
science and technology
controversy
Science and technology & public participation
37 documents
USA
Portugal
France
Italy
Singapore
Environnemental impacts & land use changes
9 documents
USA
Paris
Brazil
Europe
Northern and Southern Americas
Ecosystem services and farmers
16 documents
France
Hungary
Europe
Romania
Estonia
Technological cooperation & technological competition
5 documents
Indonesia
USA
Vaccine & controversy
11 documents
Paris
USA
Boston
France
United Kingdom

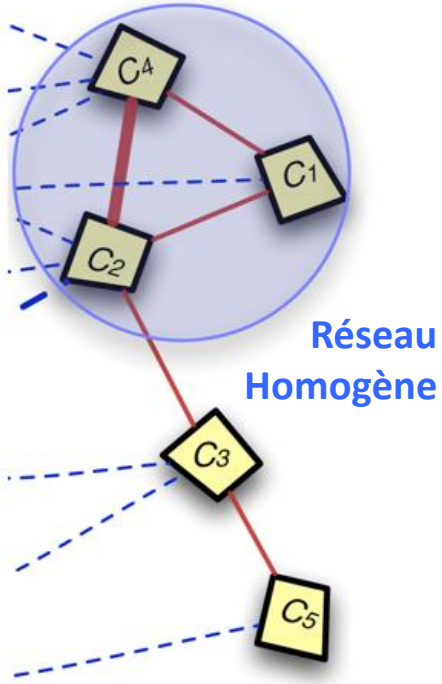
Combiner relations directes et indirectes : réseau hétérogène



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

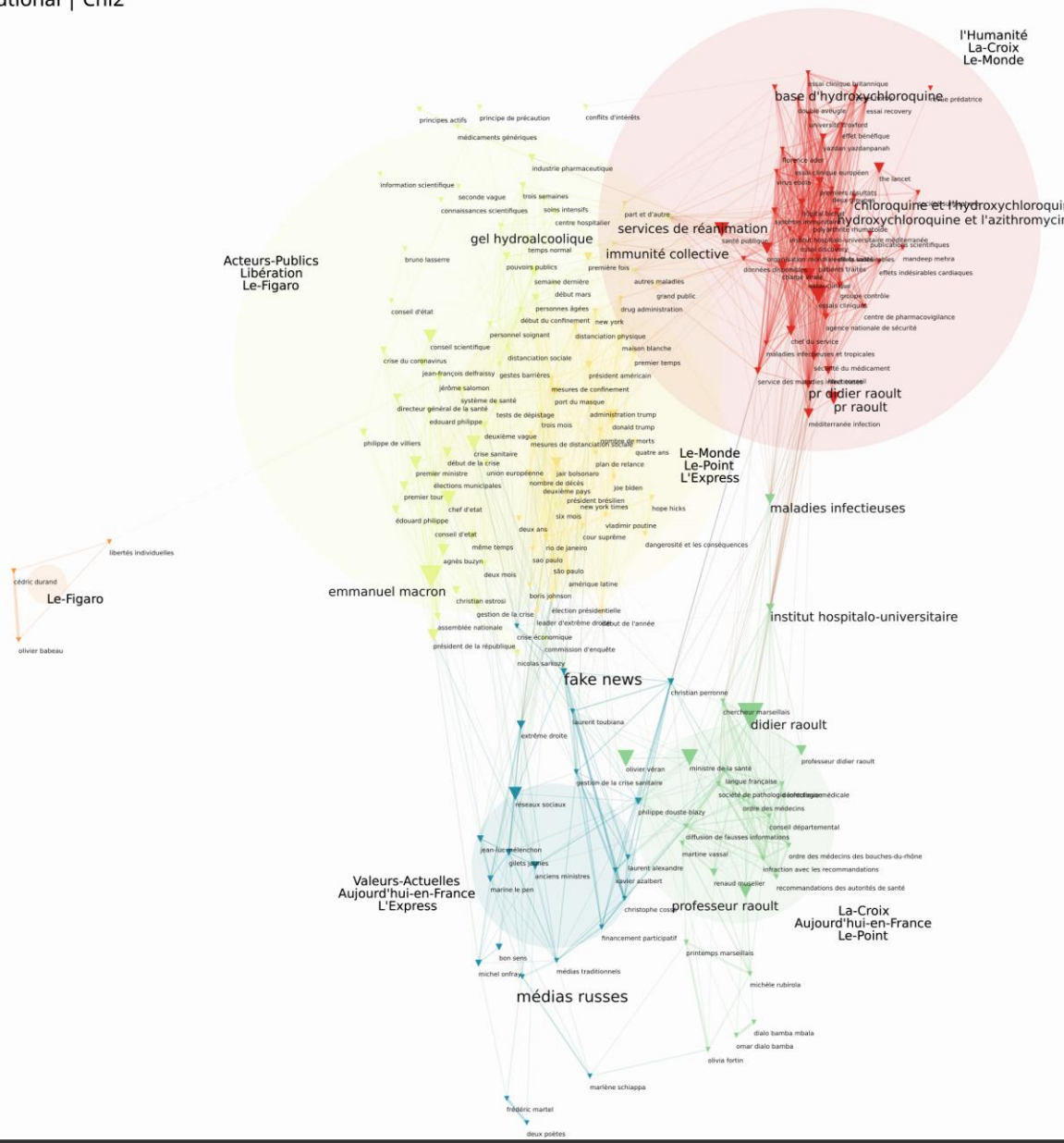
proximity measures	type of network	normalisation	special properties
raw	interaction network (*e.g.* social network)	no	-
χ^2	homogeneous & heterogeneous	yes	normalization tend to create links toward higher degree nodes
MI	homogeneous & heterogeneous	yes	Inspired from information theory
Cramer	homogeneous & 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

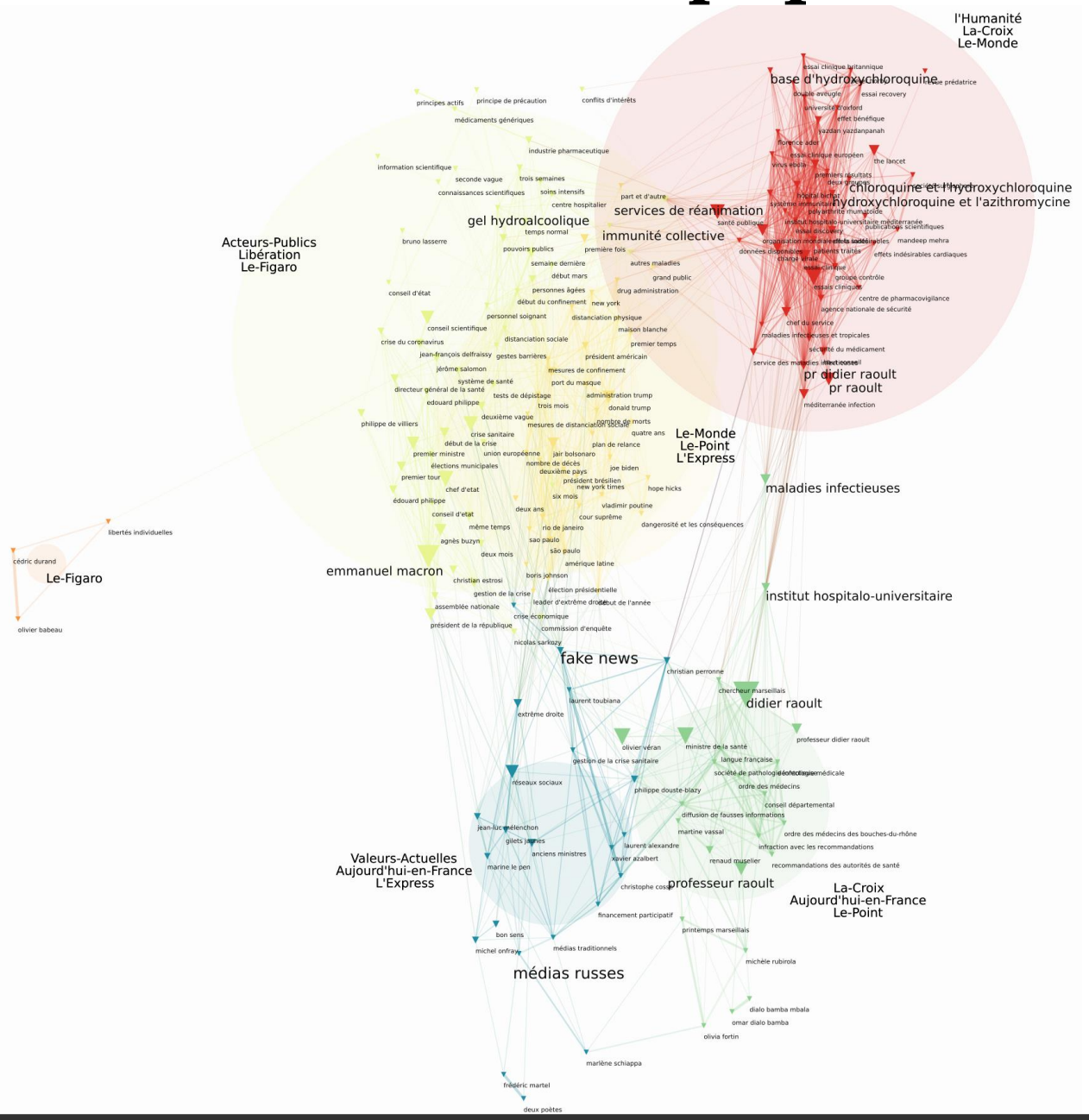


Nombre de clusters (espaces sémantiques)

Exemples de métriques

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

Lecture mésoscopique

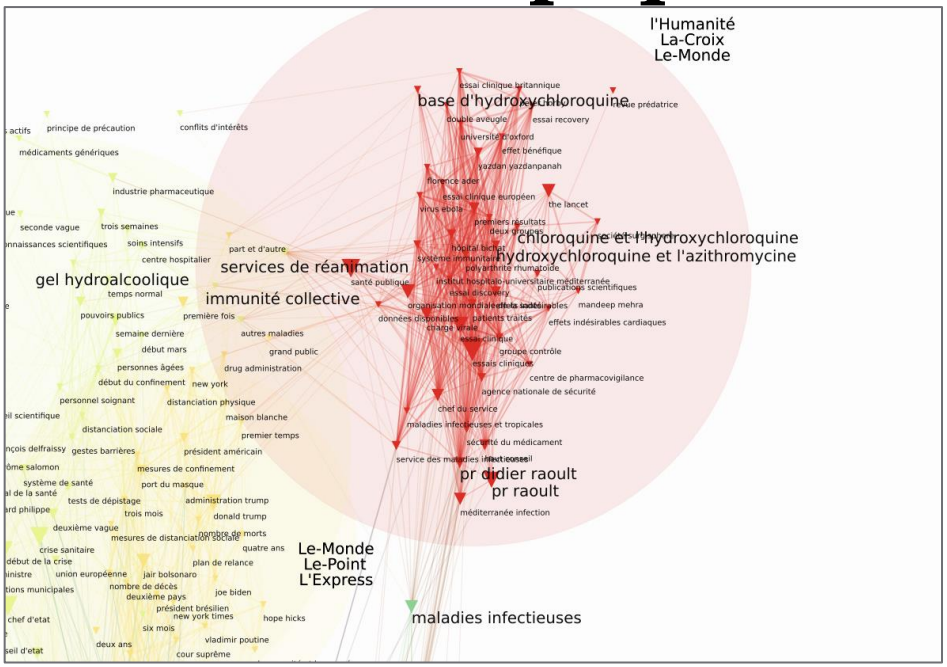


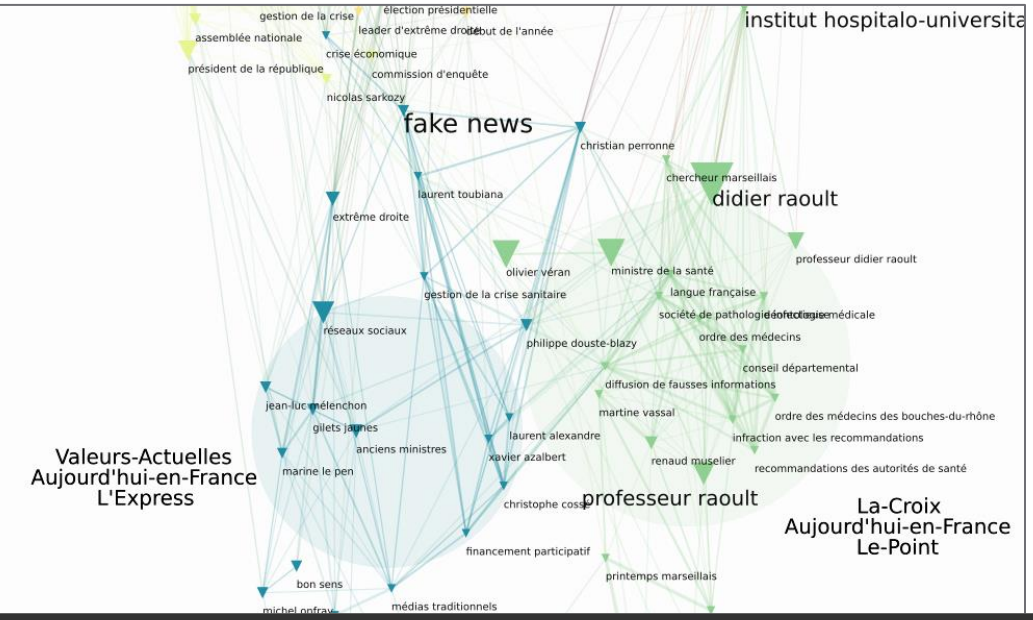
Proximité entre les clusters (espaces sémantiques interstitiels)

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





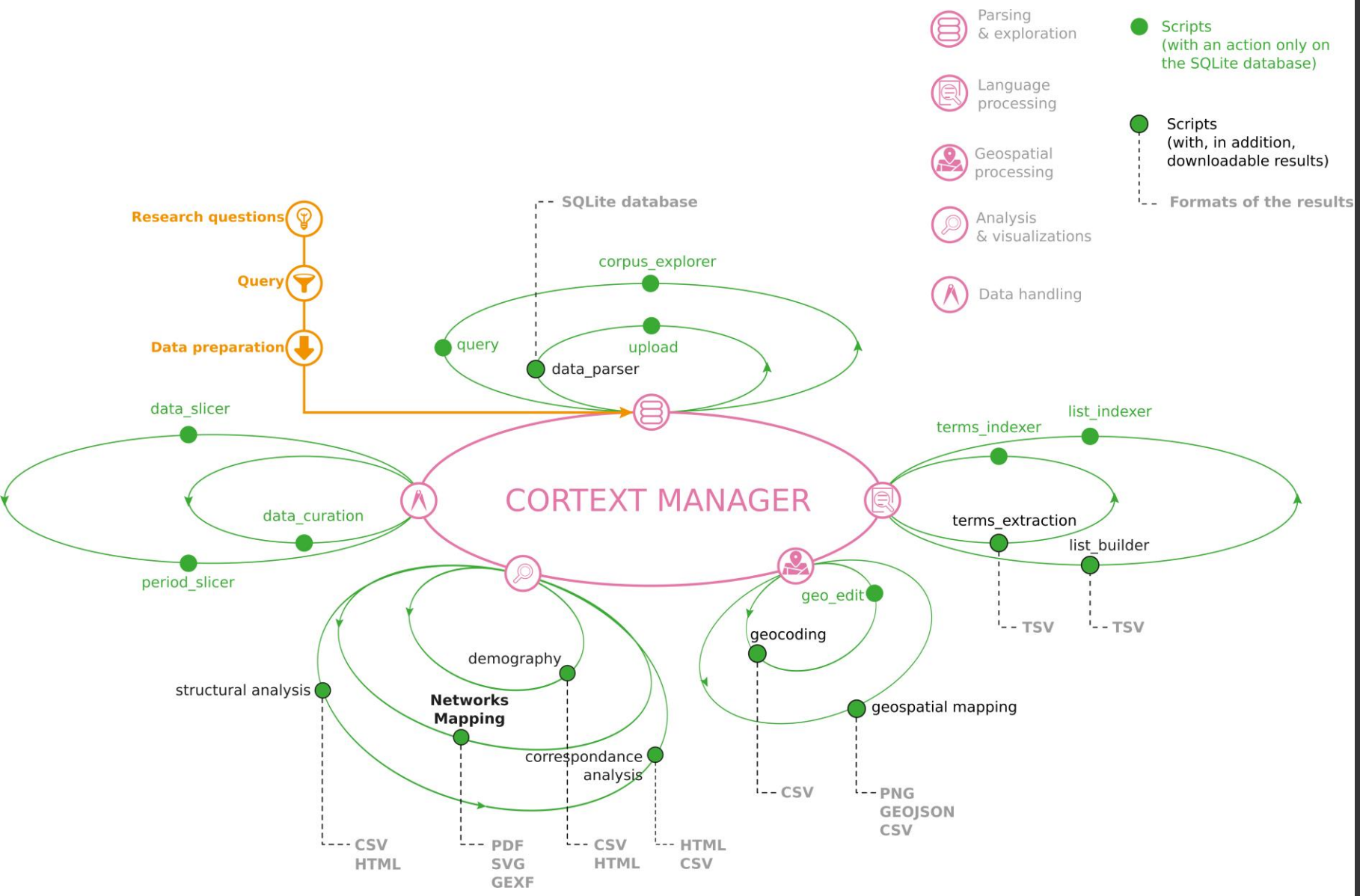
Interprétation locale de la composition des clusters

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/trainings/cortext-tutomat-2022/

Et

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