## *1/ Choisissez un corpus qui déterminera la question que vous allez vous poser*

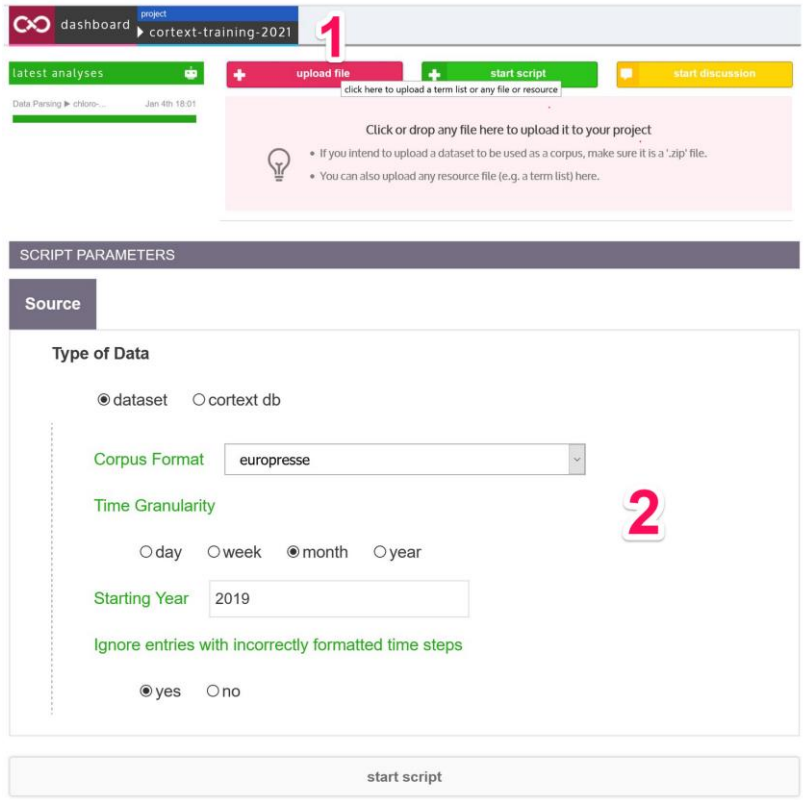
**Espace scientifique** (Web Of Science) : production académique sur la chloroquine

- Tous les articles scientifiques publiés en langue anglaise, entre décembre 2019 et novembre 2020, et accessible sur le la plateforme du Web of Science (datastet **chloro-sci-2020.zip** parser ISI sur Cortext Manager) ;
- Pour aller plus loin : tous les articles scientifiques publiés en langue anglaise, entre 2001 et 2020, et accessible sur le la plateforme du Web of Science (dataset **chloro-sci-2001-2020.zip** et parser ISI sur Cortext Manager). Ce corpus est déconseillé pour une première découverte de CorText Manager.

**Presse nationale française** (Europress)

- Tous les articles de journaux de la presse nationale française publiés entre décembre 2019 et novembre 2020, et accessible sur le la plateforme Europress (dataset **chloro-presse-fr.zip** parser europress sur Cortext Manager).

*Inscrivez-vous et créez-vous un projet sur CorText Manager : https://managerv2.cortext.net/*
*Télécharger les corpus : https://docs.cortext.net/trainings/cortext-lisis/1-introduction/01-dataset/*
*Uploadez et parser le corpus*




**ISI** pour le format les notices d'articles scientifiques (Web Of Science)
**europress** pour les articles de journaux nationaux en France : chosissez **month** pour **Time Granularity** et préciser **2019** pour **Starting year**

## *2/ Puis choisissiez une dimension d'analyse et une question*

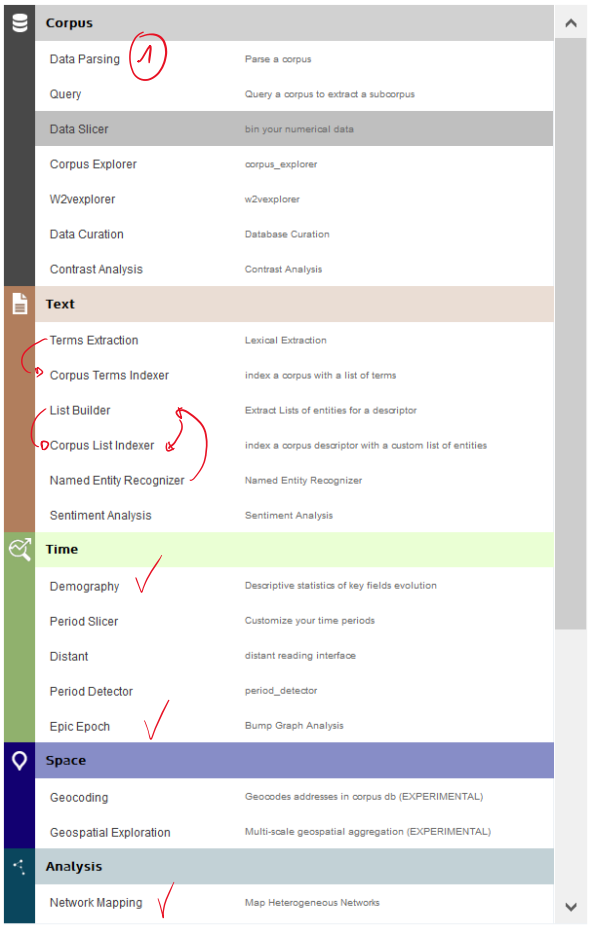
Dimensions d'analyse

- **Analyse sémantique** (réseaux de mots et identification des thèmes)
- **Analyse sociale** (réseaux de chercheurs, d'organisations, de lieux géographiques)
- Il est possible de croiser les deux (**socio-sémantique**) et d'utiliser la **dimension temporelle**

Exemples de questions, choisissiez-en une ou construisez votre propre question à partir de ces exemples

- En 2020, qu'elles ont été les **sources bibliographiques mobilisées** dans les travaux des chercheurs sur ces sujets et quelles sont les « écoles de pensées » (relations directes | Corpus WOS : Network Mapping sur la variable Cited Ref) ?
- Quelles sont les **espaces sémantiques** qui ont structurés les débats dans la presse nationale française en 2020 sur ces sujets (relations Indirectes | Corpus Europresse : lexical extraction sur la variable content) ?
- Quelles sont les **lieux géographiques** les plus souvent mentionnés dans les débats s'exprimant dans la presse nationale française en 2020 sur ces sujets (relations Indirectes | Corpus Europresse : Name Entity Recognition sur la variable content) ?
- Quelles sont les **espaces géographiques** dont les chercheurs ont été les plus actifs en 2020 sur ces sujets (relations directes | Corpus WOS : Network Mapping sur la variable cities) ?
- Quelles sont les **organisations** dont les chercheurs ont été les plus actifs en 2020 sur ces sujets, et **comment collaborent-elles** (relations directes | Corpus WOS : Network Mapping sur la variable Research institutions) ?
- …

## 3/ Utilisez les scripts suivants




- Parsing (automatique au moment de l'upload)
- Time > demography

**Travail du texte**

- Text > Term extraction + Corpus term indexer
- Text > List builder + Corpus List indexer
- Text > Name Entity Recognizer + List Builder + Corpus List indexer
  *Aidez-vous des ressources si l'une d'elles correspond à votre question* *https://docs.cortext.net/trainings/cortext-lisis/1-introduction/02-dictionnaires/*

**Analyser**

- Time > demography
- Time > Epic Epoc
- Analysis > Network heterogenous network

Dans le script **Network Mapping** il est demandé de préciser la mesure qui sera utilisée pour calculer la proximité / similarité entre deux variables (**onglet Edge** « promixity mesure » ou **l'onglet « Network Analysis and layout »** quand « Add information from a 3rd variable to tag clusters or produce a heatmap » est activé).

- *Pour aller plus loin :* https://docs.cortext.net/analysis-mapping-heterogeneous-networks/mapping/#tagging-heatmap-specificity-measure

| proximity measures | type of network | normalisation | special properties |
|---|---|---|---|
| raw | interaction network (*e.g.* social network) | no | - |
| $X^2$ | 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) |

**Raw** correspond à la valeur brute (le compte, la fréquence). Mesure de cooccurrence brute. Par exemple, on comptera 1 pour la paire {carottes, poireaux} à chaque fois que carottes et poireaux apparaîtront ensemble dans une recette. Mesure pertinente pour la construction de réseaux de collaborations (aucune correction particulière de l'information est nécessaire ; respect des données). Repose sur l'hypothèse qu'un lien correspond à une interaction effective.

- est généralement à privilège pour **les réseaux sociaux** (collaborations entre des individus ou entre des organisations)
- *pour aller plus loin : https://docs.cortext.net/metrics-definitions/#raw*

**Distributional** : issue de la linguistique récente, elle permet de faire apparaître des relations pertinentes, bien que rares. La proximité distributional s'appuie sur la mesure directe d'Information Mutuelle précédemment présentée. Pour un mot donné, l'Information Mutuelle est la quantité d'information apportée par la présence de ce mot dans le contexte d'apparition d'un autre mot. La mesure Distributional, pour deux termes (i et j), compare donc les vecteurs à n dimensions d'Information Mutuelle de ces deux termes, autrement dit la similarité des contextes d'apparition de ces termes. Cela permet de détecter des synonymes, c'est-à-dire des termes qui ne cooccurrent pas forcement mais qui ont des contextes d'apparition identiques. Elle a donc une propriété d'interchangeabilité : carotte et porreau étant des légumes, ils ont des caractéristiques communes, des possibilités d'associations avec d'autres ingrédients proches ainsi que des modalités de cuissons similaires.

- est généralement à privilège pour **l'analyse sémantique** : très performant pour extraire la structures sous-jacentes des textes, en présentant les mots qui jouent des fonctions similaires dans les textes
- *reference: https://docs.cortext.net/metrics-definitions/#distributional*

**Chi²** : mesure l'intensité du lien entre deux termes en appréciant l'écart par rapport à la valeur attendue. Le Chi²est donc mesure de spécificité. La valeur attendue entre deux termes est égale à la somme de l'ensemble des cooccurrences du premier terme (avec, donc, l'ensemble des autres termes) multipliée par la somme de l'ensemble des cooccurrences du second terme (avec, donc, l'ensemble des autres termes), sur la somme de l'ensemble des cooccurrences entre elles (somme des lignes plus la somme des colonnes de la matrice de cooccurrences, soit, en fait, le nombre total de cooccurrences observées). Lorsqu'il est positif, l'écart entre la valeur réelle des cooccurrences de deux termes et entre la valeur attendue indique une surreprésentation du lien entre ces deux termes et donc une **spécificité**.

- est généralement à privilège pour **dégager la spécificité d'une valeur** dans un contexte (par exemple avec **l'onglet « Network Analysis and layout »** quand « Add information from a 3rd variable to tag clusters or produce a heatmap » est activé)
- *reference : https://docs.cortext.net/metrics-definitions/#chi2*

## 4/ Protocole méthodologique et résultats

- Reportez dans un document vos étapes, et ajouter vos résultats.
- Comparez avec les propositions de résultats : *https://docs.cortext.net/trainings/cortext-lisis/1-introduction/03-resultats/*