CORLI : Un corpus ouvert du français
– ou comment travailler à rassembler les briques existantes ?

Personnes impliquées dans le projet

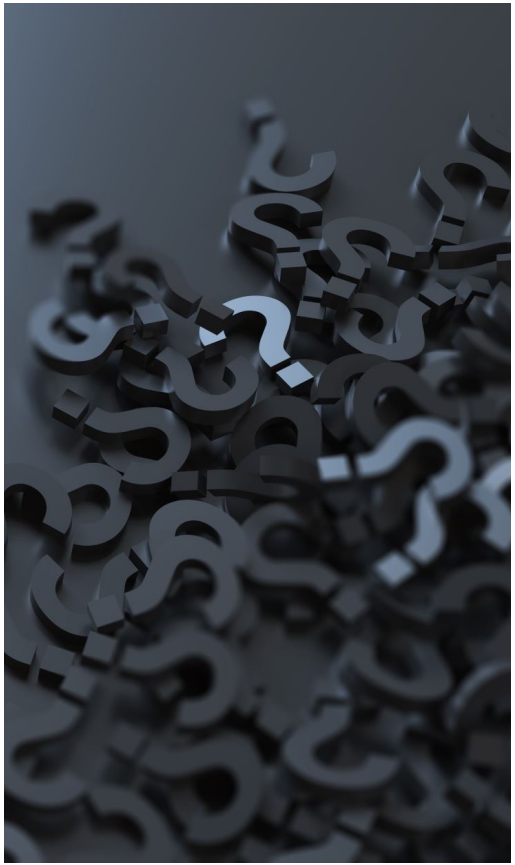
- **10 membres du comité de pilotage de Corli:**
- **Christophe Parisse, Céline Poudat, Flora Badin, Christophe Benzitoun, Sascha Diwersy, Carole Etienne, Julie Glikman, Marie-Paule Jacques, Amalia Todirascu, Agnès Tutin**
- **Notre ingénieure d'étude CORLI: Mathilde Guernut**
- **Le consortium ARIANE pour les corpus de littérature: Fatiha Idmhand, Ioana Galleron, Sabine Loudcher, Alexei Lavrentev, Geoffrey Williams**

Un projet du consortium

- **Objectif général**:
  - une ressource représentative et volumineuse
  - une ressource librement accessible
  - une ressource organisée
  - une ressource standardisée
  - une ressource de qualité et outillée
- … pour le français à partir des corpus existants
- Et pour cela, récupérer, éditer, compléter, les corpus et métadonnées existants.
- Se rapprocher des autres données de langage existantes, comme par exemple les données éditoriales (notamment littéraires) sur lesquelles s'est spécialisé Cahier/Ariane

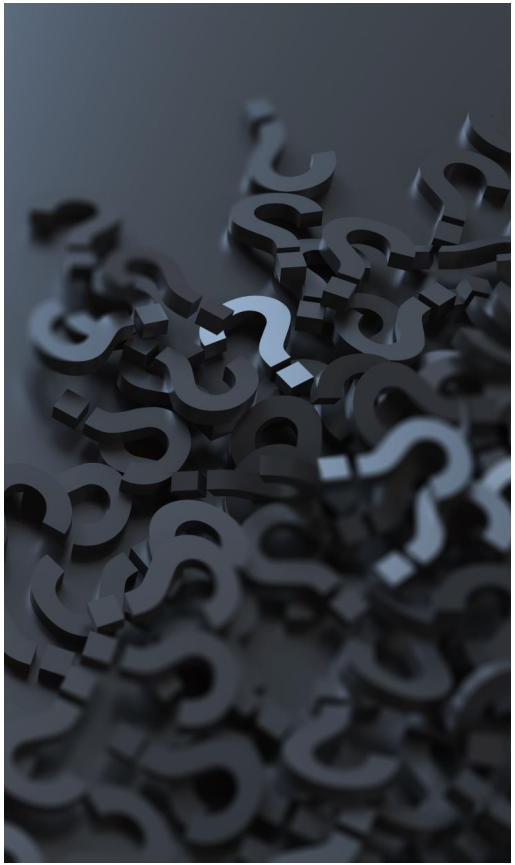
Une très grande variété de métadonnées, de format et de contenus

- **Cette variété a de grandes conséquences sur les outils que l'on peut associer aux données et aux contenus**
  - **Les métadonnées pour organiser l'interrogation des données**
  - **Les données pour accéder aux contenus**
  - **Les contenus pour avoir des outils adaptés (écrit, oral, CMC, multimodal)**



Quel est l'état des métadonnées en linguistique?

- **Les métadonnées sont une évidence quasi "intuitive" pour un linguiste.**
- **Un travail de recherche en linguistique nécessite de savoir:**
  - **quelle langue, quel dialecte, quel argot, … on parle ?**
  - **qui parle, qui écrit, où, quand, comment ?**
- **On trouve donc beaucoup de métadonnées mais leur format est prévu pour être lu par un humain et non par une machine.**



Quel est l'état des métadonnées en linguistique?

- **On peut distinguer plusieurs types de métadonnées, par exemple:**
  - **Les métadonnées documentaires (date de publication, éditeur…)**
  - **Les métadonnées qui joueront un role dans l'analyse (hypothèse de contraste) pour le chercheur**
  - **Les métadonnées de recherche (*end-user*) plus intuitives (*e.g.* thèmes)**

Se baser sur les corpus existants et validés par la communauté

- **Les corpus disponibles sur les plateformes de dépôt, conservation et diffusion:**
- **Ortolang, Cocoon, Nakala, Autres (CHILDES, dépôts spécifiques)**

- **Pour toutes les plateformes, deux options (complémentaires) sont envisagées:**
  - **Sélectionner les corpus selon des critères de qualité à déterminer par la communauté**
  - **Accéder à tous les corpus qui sont FAIR**

Plan de travail

Récupération, nettoyage et harmonisation des métadonnées

Pour les éléments de métadonnées, se baser sur un travail commun Ariane/Corli pour les corpus écrits, sur le travail du consortium CORLI 1 pour les corpus oraux, sur le projet CoMeRe pour les métadonnées de la CMC

Récupération des formats, conversion vers un format TEI

Trois sous-versions de la TEI envisagées: Tags <p> pour l'écrit, Tags <u> pour l'oral et le multimodal, Tags <post> pour la CMC

Mettre en place des outils d'interrogation:

Interrogation en ligne au format TEI : ALLEGRO (moteur de Frantext)

Interrogation locale au format TEI après téléchargement: TXM

Interrogation au format texte: Content Search de CLARIN

En 2023-2024 focalisation sur les corpus déposés sur ORTOLANG

Permet de tester en situation réelle les chaînes de traitement dont on dispose:

- **TEICORPO pour les conversions depuis des formats de l'oral**
- **TEIMETA pour l'édition complémentaire des métadonnées**
- **Moteur ALLEGRO de Frantext pour l'interrogation des données**

Trois stages sont en cours pour réaliser:

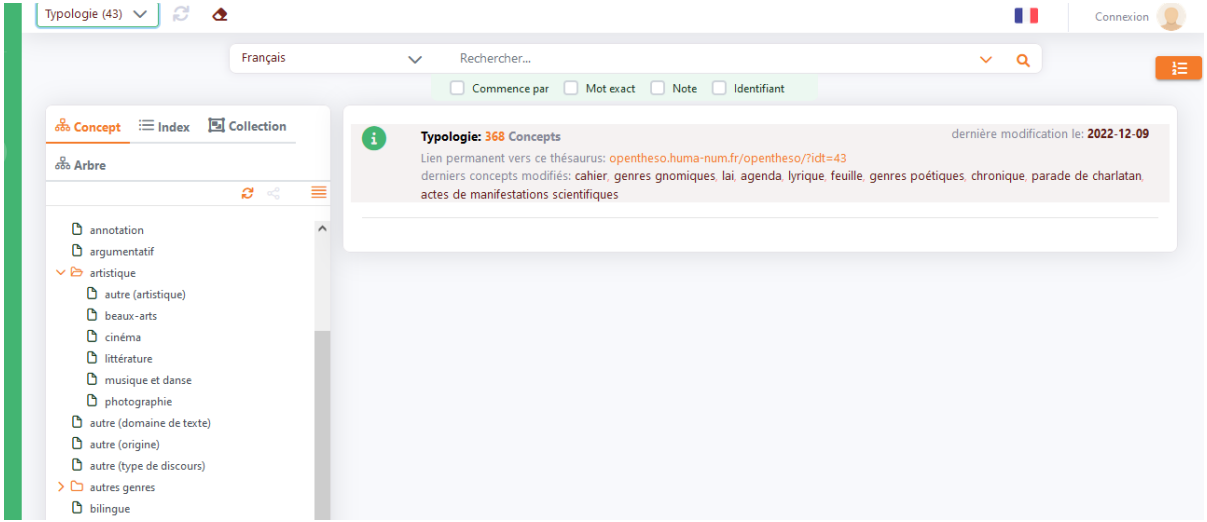
- **l'interface système d'ALLEGRO**
- **l'interface d'interrogation d'ALLEGRO**
- **le contrôle et la conversion des formats de données déposées sur ORTOLANG**

Détail du travail sur les métadonnées

- **Evaluation des corpus déposés sur Ortolang et de leurs métadonnées**
- **Détermination d'un header TEI minimal**
- **Harmonisation des headers TEI**
  - **Projeter ou étendre les métadonnées pour**
    - **avoir des métadonnées au niveau corpus …**
    - **… comme au niveau des données élémentaires (textes, enregistrements, etc.)**
  - **Traitement automatique autant que possible**
  - **Traitement semi-automatique pour compléter, corriger, valider les métadonnées**

Détail du travail sur les métadonnées

- **Harmonisation des genres (collab. Cahier/Ariane - mise en place d'un vocabulaire contrôlé partagé)**
- **Harmonisation des informations en fonction des domaines (par exemple auteur/créateur/…)**



Thésaurus de genres commun avec le consortium ARIANE

Résultats attendus

- **Inclusion d'une série de corpus la plus large possible au-delà d'ORTOLANG**

- **Corpus écrits**
  - **Scientext https://scientext.hypotheses.org/corpus**
  - **Scienquest https://corpora.aiakide.net/**
  - **Archives parlementaires https://archives-parlementaires.persee.fr/**
  - **Consortium CAHIER https://cahier.hypotheses.org/**
  - **E-CALM https://www.ortolang.fr/market/corpora/e-calm**
  - **Corpus 14 https://www.univ-montp3.fr/corpus14/**
  - **Democrat https://hdl.handle.net/11403/democrat**

- **Corpus oraux**
  - CFPP2000 http://cfpp2000.univ-paris3.fr/search.html
  - ESLO http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus
  - CHILDES https://talkbank.org/DB/
  - CT3-ORTOLANG https://ct3xq.ortolang.fr/ct3xq/interro
  - PFC https://public.projet-pfc.net/transcription/

- **Corpus multiformats**
  - CEFC-Orféo https://orfeo.ortolang.fr/, http://orfeo.grew.fr/
  - CoMeRe http://hdl.handle.net/11403/comere

- **Autres types d'initiatives, corpus semi-ouverts**
  - Partie libre de Frantext
  - Les bases d'Hyperbase (E. Brunet) http://ancilla.unice.fr/pages/bases/
  - Bibliothèque Nationale de France