VERS UN CORPUS OPTIMAL POUR LA FOUILLE DE TEXTES

L'EXEMPLE DE BEETHOVEN

Journée ISTEX - 19 Janvier 2021

Camille de Salabert | Inist-CNRS | 2021

UN CORPUS OPTIMAL POUR LE TDM ?

Méthodologie de Constitution de Corpus à partir d'ISTEX

Littérature

- ★ peu d'études sur la constitution des corpus (exception des linguistes)
- ★ peu de détails sur les étapes de traitements nécessaires
- ★ connaissances informatiques requises

CHOIX DE LA THÉMATIQUE

Musique !

- ★ Un sujet adapté à l'objectif visé (exploiter la plateforme ISTEX et ses outils)
- ★ Des applications TDM possibles
- ★ Un corpus en Sciences Humaines & Sociales à ajouter sur data.istex
- ★ 2020 : 250e anniversaire de la naissance du compositeur allemand Beethoven

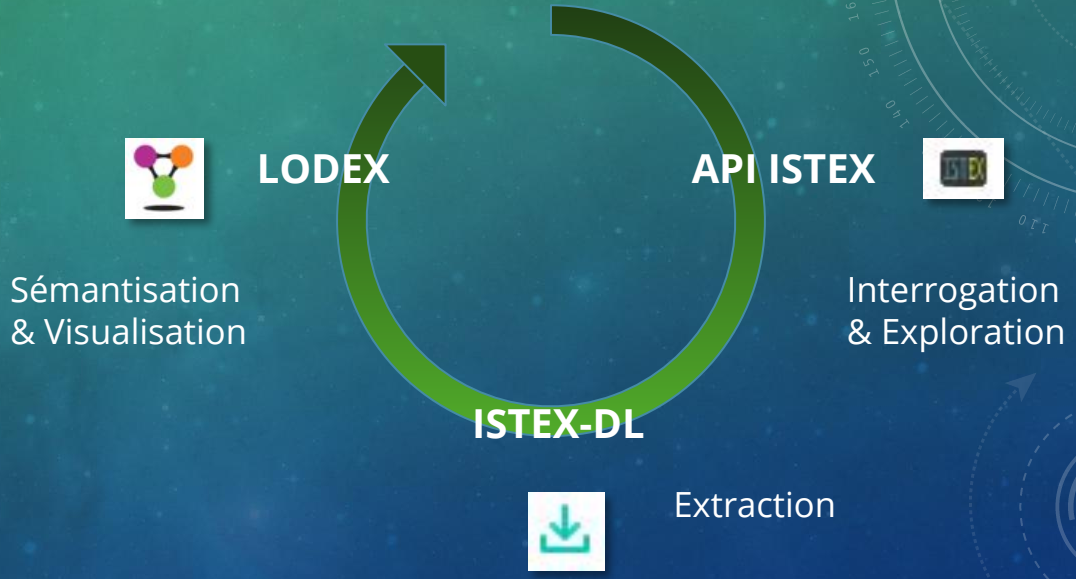
OBJECTIF

Un Corpus Beethoven
Annoté et Aligné

- ★ Créer un corpus de musicologie sur Beethoven
- ★ Détecter les entités nommées
- ★ Aligner les entités spécifiques de la musique avec DOREMUS

CONCEVOIR SON CORPUS POUR LIMITER LES TRAITEMENTS

Stratégie Itérative :
3 Outils



CONCEVOIR SON CORPUS POUR LIMITER LES TRAITEMENTS

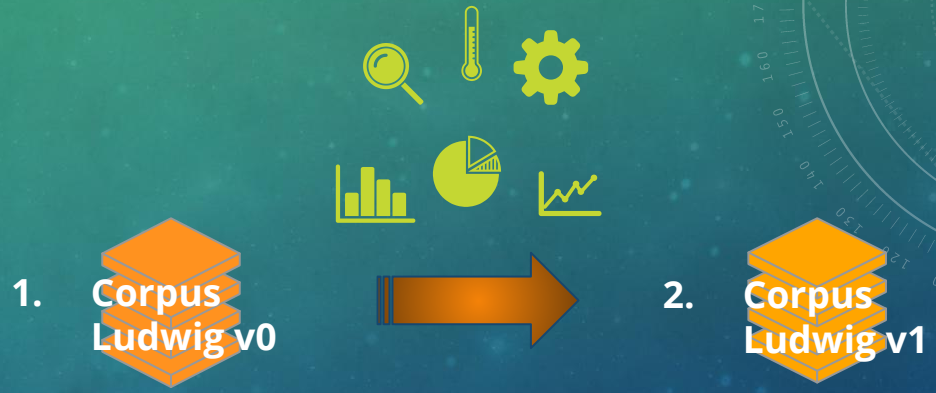
Stratégie Itérative :
2 Étapes

1. Contenu scientifique
   - pertinence
   - réduction du bruit
   - réduction du silence

2. Exploitation TDM
   - articles multiples (page ou PDF unique)
   - PDF image
   - nombre mots/PDF
   - présence de résumés
   - langue, etc.

CONCEVOIR SON CORPUS POUR LIMITER LES TRAITEMENTS

Stratégie Itérative : 2 Versions du Corpus



Collection Beethoven :
https://beethoven-collection.corpus.istex.fr/

DEMO

Liens utiles :

https://www.istex.fr/
https://data.istex.fr/
https://demo.istex.fr/
https://dl.istex.fr/
https://lodex.inist.fr/

CONGRÈS JEP-TALN-RECITAL 2020

Article dans HAL

"Vers un corpus optimal pour la fouille de textes : Stratégie de constitution de corpus spécialisés à partir d'ISTEX"

Camille de Salabert, Sabine Barreaux

https://hal.archives-ouvertes.fr/INIST/hal-02768520v3

Vidéo (fichier annexe) dans HAL

ISTEX COMME RESSOURCE TEXTUELLE

Une Archive Vivante & Outillée

- ★ Plus de 23 millions de documents de toutes disciplines
- ★ Des enrichissements générés par des outils de TDM
- ★ Différents formats homogénéisés de texte intégral et de métadonnées
- ★ Un travail sur la qualité des textes et des métadonnées
- ★ Des indicateurs et des critères de sélection
- ★ De nouveaux enrichissements prometteurs
- ★ Des outils associés & connectés
- ★ Une ressource à disposition de l'ESR pour la fouille de textes

"Ecouter Beethoven autrement avec ISTEX"

Merci !
Des questions ?