Explorer et analyser un corpus avec Lodex

Mathilde Huguin
Ingénieure de recherche Inist-CNRS, équipe *Istex - Texte & Corpus*
mathilde.huguin@inist.fr

Séminaire ONuSeL - 9 avril 2026 - *LT2D - Cergy Paris Université*

Inist cnrs

Plan

01
Introduction

02
Fonctionnement

03
Démonstration

04
Conclusion

Introduction

Présentation de Lodex

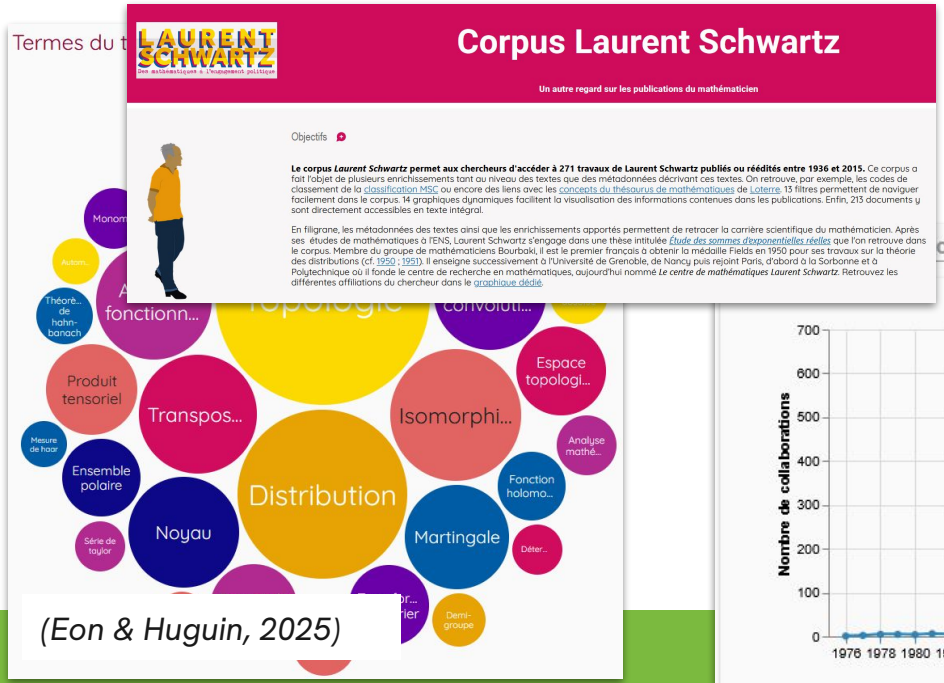
Introduction
- Éléments généraux

- Acronyme pour *Linked Open Data Experiment*
- Logiciel open source **de visualisation et d'analyse de données structurées** développé par l'Inist-CNRS (code disponible sur GitHub)
- Créé initialement en 2016 pour valoriser les données **Istex** (Gregorio *et al.*, 2019 ; Huguin & Barreaux, 2023 ; Huguin, 2025)
- Logiciel accessible en ligne, données hébergées sur les serveurs de l'Inist-CNRS (demande en ligne)
- Réservé aux ayants-droit Istex
- Version actuelle 16.10.11

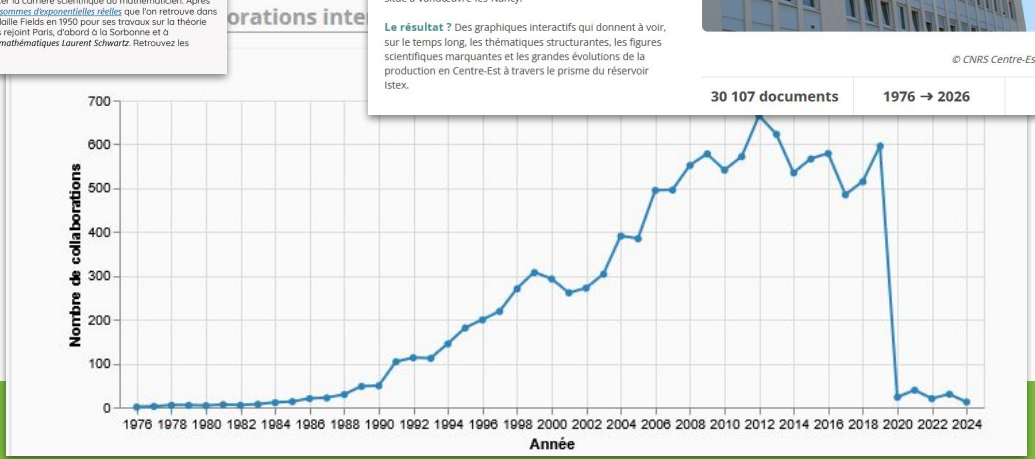
Introduction
- À quoi sert Lodex ?

- Permet de transformer un jeu de données (xml, json, csv, tsv…) en **site web dynamique**
- Offre différents angles de vue à l'aide :
  - de **graphiques** : histogramme, diagramme circulaire, cartographie…
  - de **fiches descriptives** : présentation des documents du corpus
  - de **facettes**
- Facilite la **curation et l'enrichissement** des données grâce à un catalogue de web services de TDM (*text and data mining*)

Introduction
- Exemples



(Eon & Huguin, 2025)



Fonctionnement

Comment utiliser Lodex ?

Fonctionnement
- Concept de base

Données

L'import d'un jeu de données ou d'un corpus (xml, json, csv, tsv, tei...) est la première étape vers la publication avec Lodex.

Modèle

Le modèle rassemble toutes les instructions décrivant la manière dont les données vont être mises en forme à des fins de visualisation.

Site web

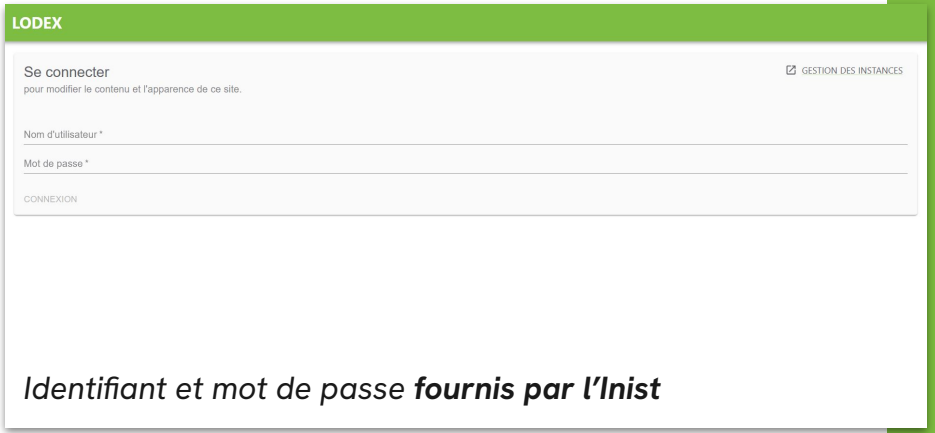
Une fois les données chargées et le modèle appliqué le site peut être publié.

Fonctionnement
- Concept de base

Fonctionnement - Connexion

3 rôles possibles lors de la connexion

- *Administrateur* : peut modifier le contenu et l'apparence du site
- *Utilisateur* : peut consulter le site (rôle optionnel : données sensibles, travail en cours, etc.)
- *Contributeur* : peut annoter les différents éléments du site (rôle optionnel : travail collaboratif, révisions, etc.)



Identifiant et mot de passe ***fournis par l'Inist***

3 onglets principaux

Menu : exporter ou supprimer des données, gérer les rôles, configurer le thème général

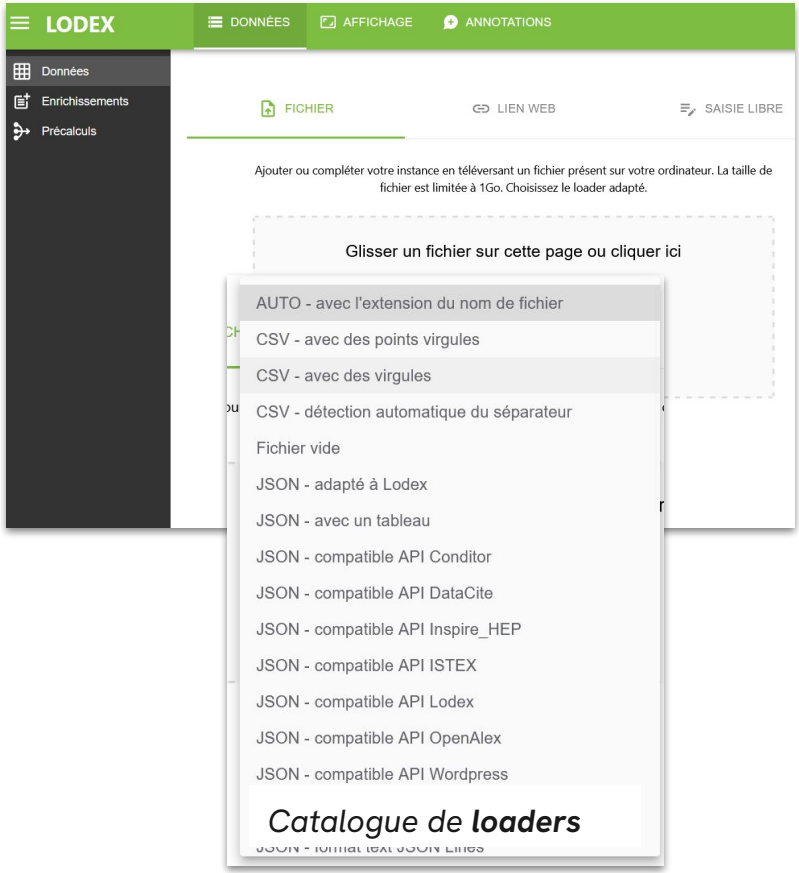
*Interface administrateur, onglet **Données**, avant l'import du corpus*

Fonctionnement - Import des données

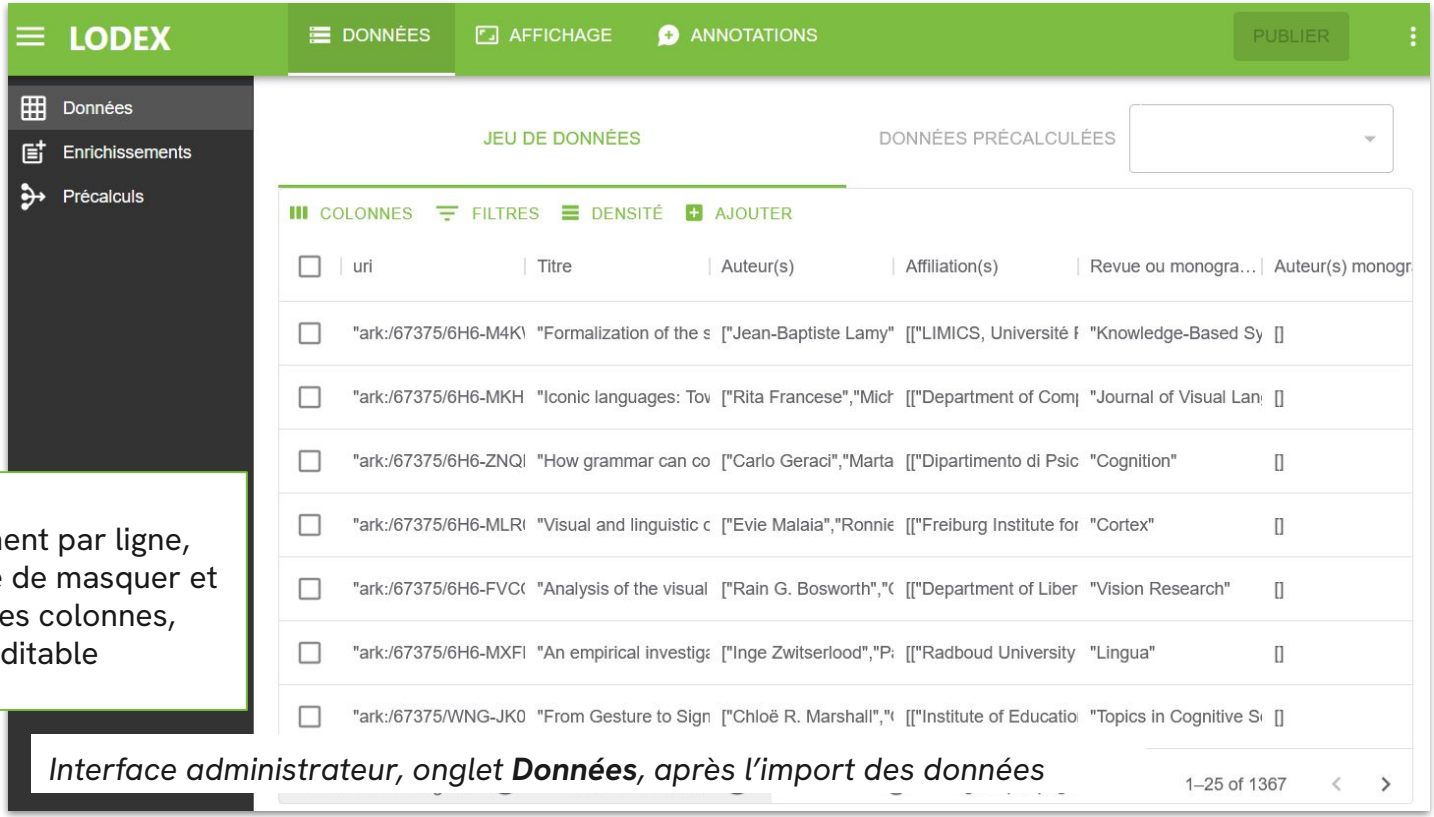
2 étapes

1. Ajouter des données (xml, json, csv, tsv...)
   - via un fichier en **local**
   - via un fichier situé sur un serveur distant (**URL**)
   - via une base externe (**API**)
   - manuellement

2. Choisir un **loader** (script de chargement des données) adapté au format des données



*Catalogue de **loaders***



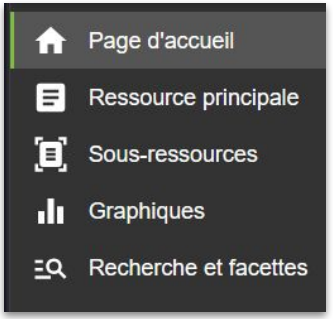
Un document par ligne, possibilité de masquer et de filtrer les colonnes, contenu éditable

*Interface administrateur, onglet **Données**, après l'import des données*

Il est possible de créer un **modèle** de A à Z ou d'utiliser un modèle existant.

Interface administrateur, onglet **Affichage**, *après l'import des données*

Fonctionnement
- Import du modèle

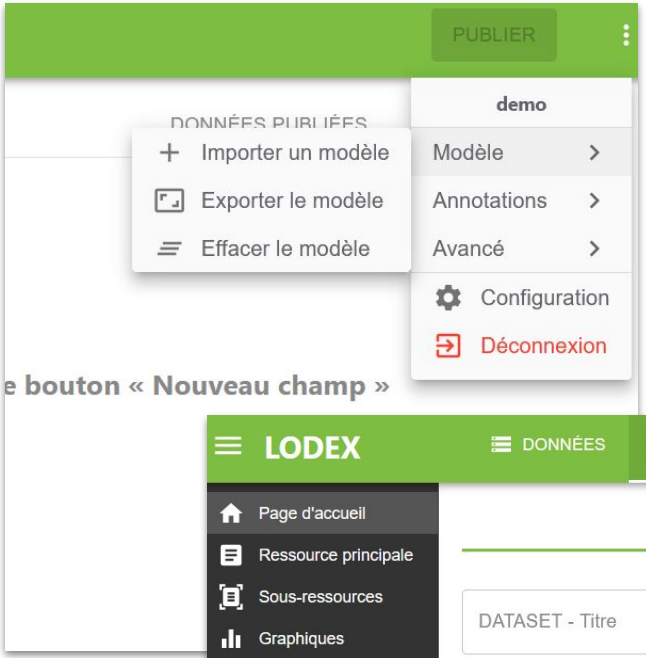


Le modèle permet de **paramétrer et personnaliser le site web** dans 5 sous-onglets

- La page d'accueil
- Les pages contenant les documents (*Ressource principale*)
- Les pages transversales (*Sous-ressources*)
- Les graphiques (*Graphiques*)
- Les éléments de recherche et les filtres (*Recherche et facettes*)

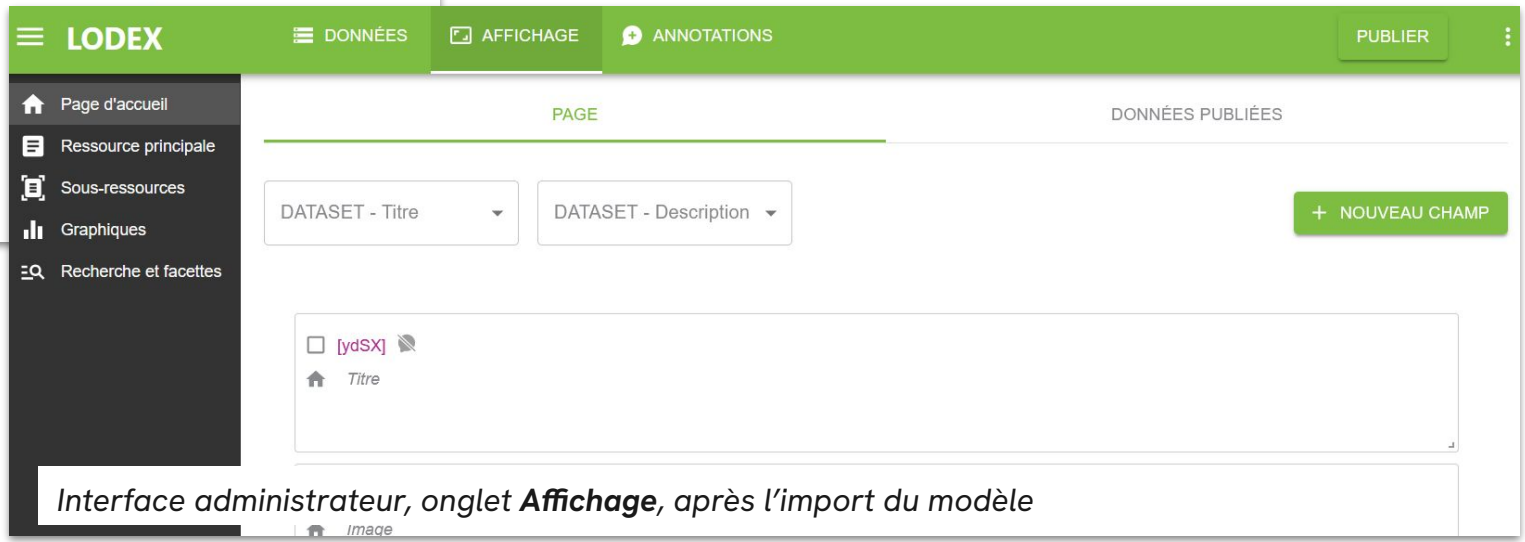
Le modèle permet d'**effectuer des opérations sur les données**

- À l'aide d'*opérations de transformation* (scripts intégrés) : remplacer une chaîne de caractères, masquer une valeur, dédupliquer…
- À l'aide de traitements avancés et personnalisés en Lodash (bibliothèque JavaScript)
- À l'aide de web services de fouille de textes



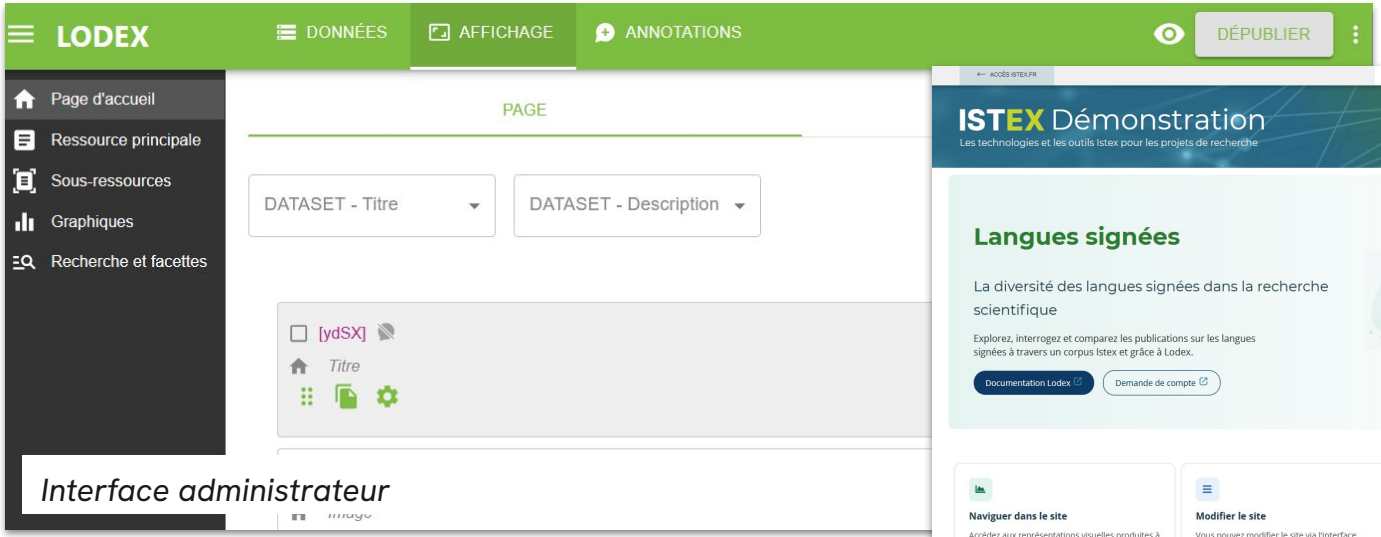
Le modèle est un fichier json ou tar.

Le site peut être publié.



*Interface administrateur, onglet **Affichage**, après l'import du modèle*

Fonctionnement - Publication



Interface administrateur

Site publié

Fonctionnement - Navigation

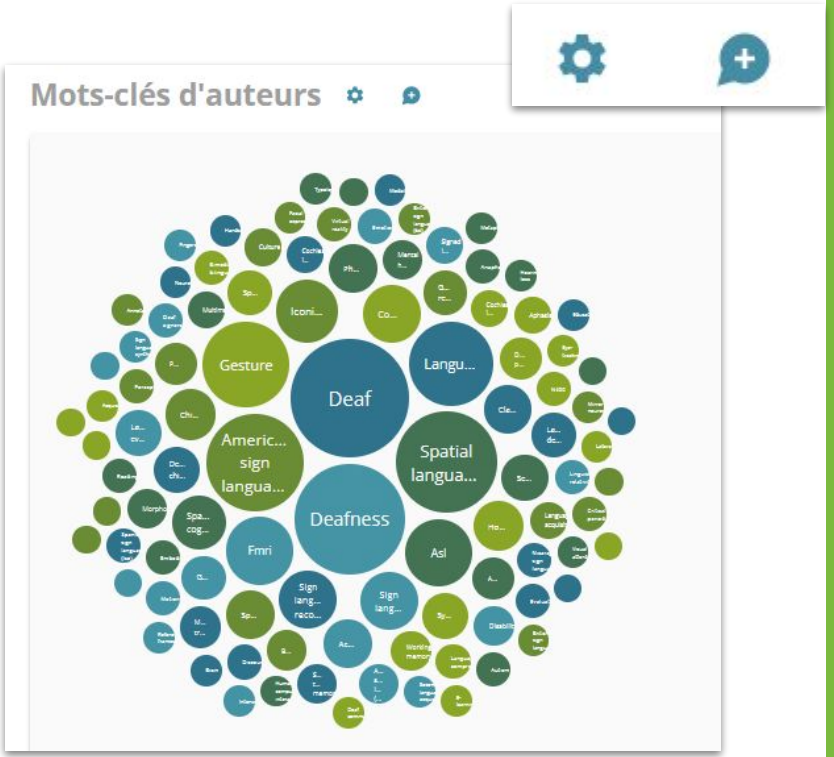
Les sites Lodex comportent actuellement 4 onglets

- ***Accueil*** : page d'accueil
- ***Graphiques*** : accès aux graphiques
- ***Recherche*** : accès aux documents du corpus
- ***Voir plus*** : connexion et accès à l'interface administrateur

Fonctionnement - Navigation

En tant qu'administrateur, les champs affichent 2 logos permettant

- D'accéder au paramétrage du champ dans l'interface administrateur
- D'annoter le champ



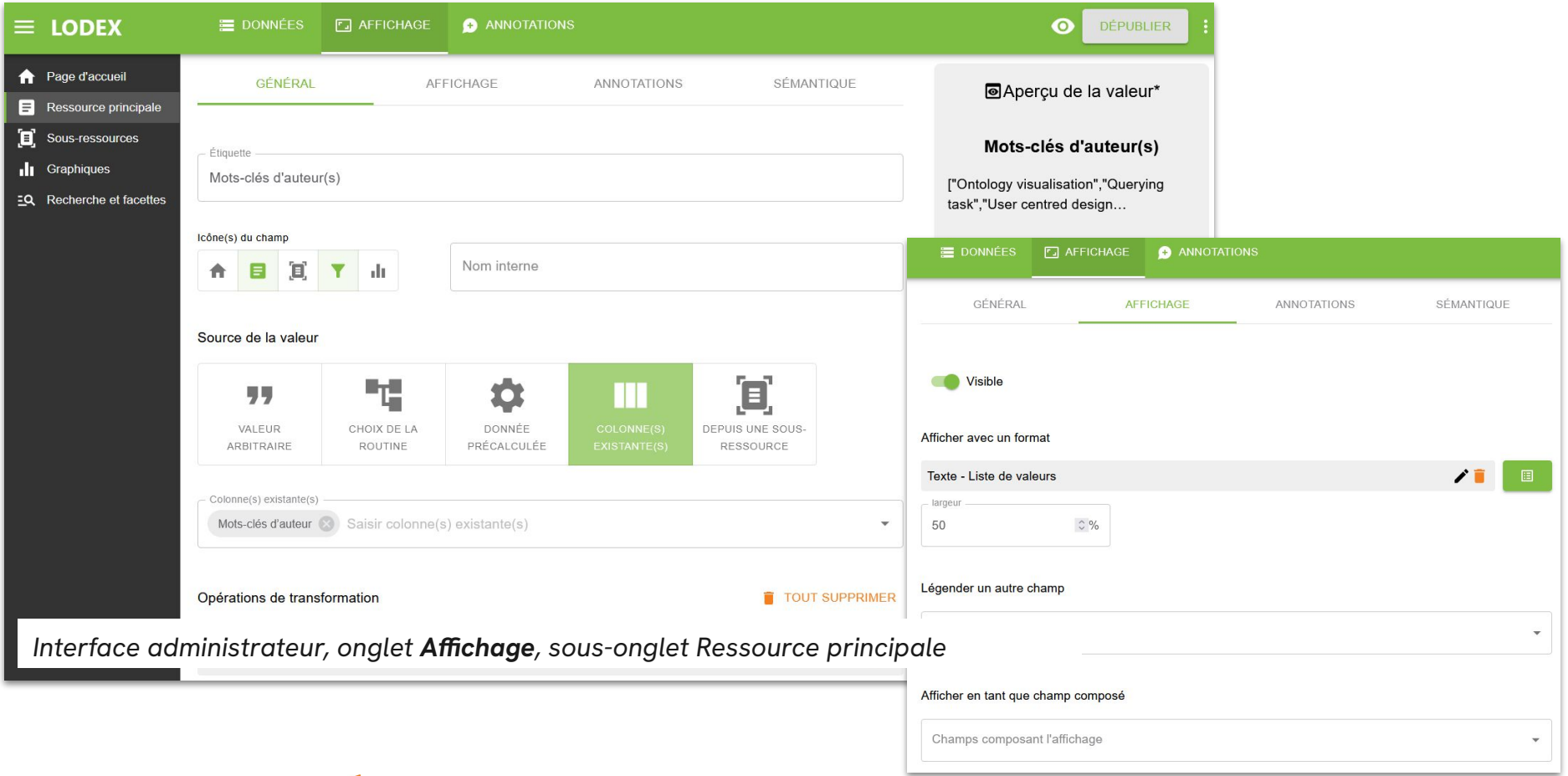
Fonctionnement
- Création d'un champ

Création d'un champ* relatif à un document

- Nommage du champ (*Étiquette*)
- Création d'une valeur arbitraire ou récupération des données depuis une colonne du corpus
- Choix du format d'affichage

Création d'un graphique

- Nommage du champ (*Étiquette*)
- Choix d'une *routine* (script qui peut effectuer des agrégations, des calculs, des reformatages)
- Récupération des données depuis un champ*
- Choix du format d'affichage (catalogue de graphiques intégrés)

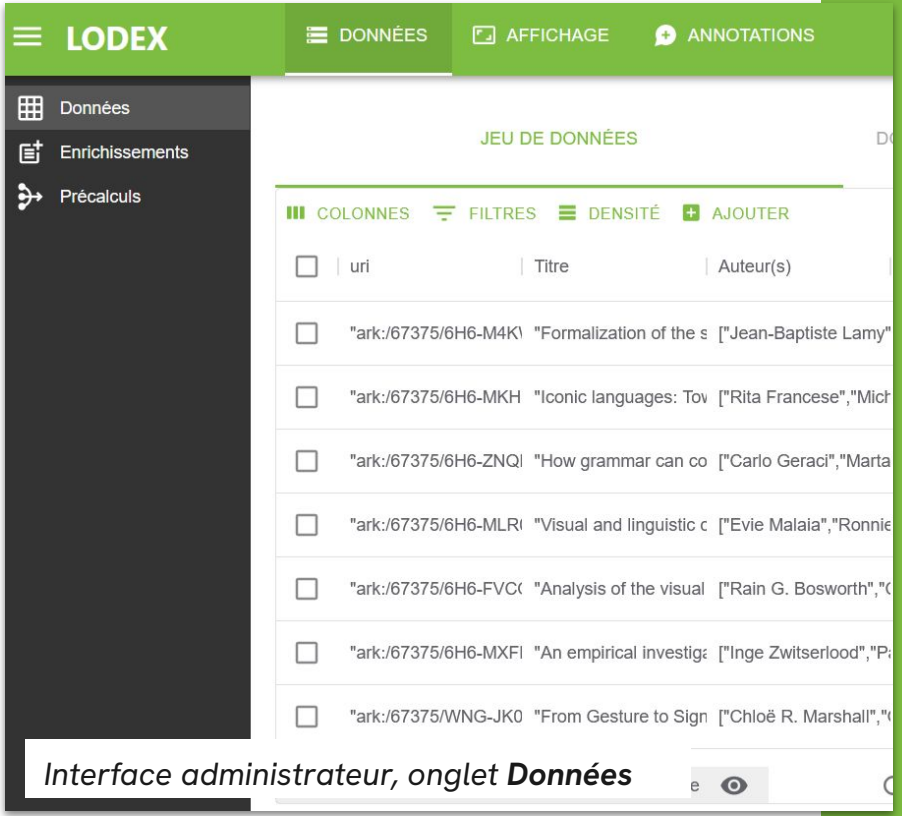


*Interface administrateur, onglet **Affichage**, sous-onglet Ressource principale*

Fonctionnement - Enrichissements

L'onglet *Données* donne accès à 3 sous-onglets

- ***Données*** : données chargées et résultats des enrichissements et des précalculs
- ***Enrichissements*** : tableau de création et de lancement des enrichissements
- ***Précalculs*** : tableau de création et de lancement des précalculs

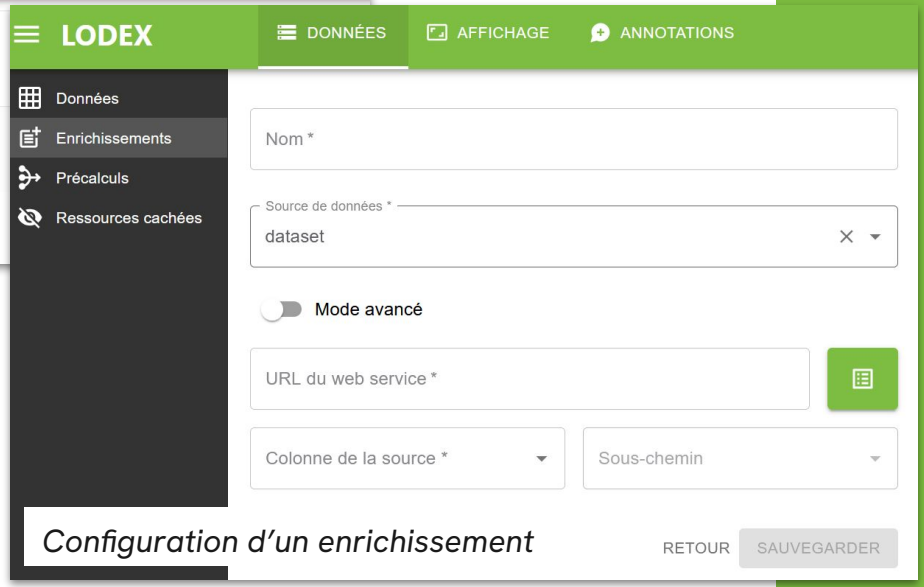


*Interface administrateur, onglet **Données***

Fonctionnement

- Enrichissements

*Interface administrateur, onglet **Données > Enrichissements***



Configuration d'un enrichissement

Fonctionnement
- Enrichissements

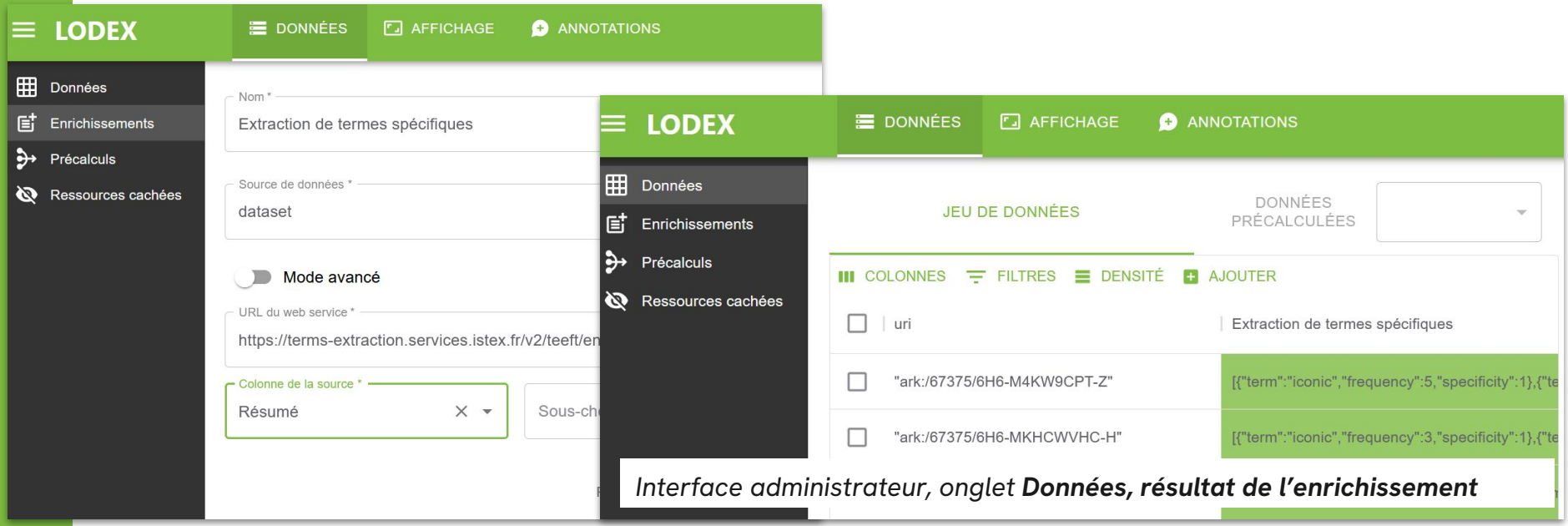
Un enrichissement est un traitement appliqué sur les données du corpus : soit via la **sélection d'un web service de fouille de textes dans le calalogue** soit en codant le script en Lodash.

Istex met a disposition **49 web services** dans Istex TDM (Cuxac, 2024)

- Un web service est un programme mono-tâche, frugal avec un paramétrage minimal
- Les web services peuvent être utilisés dans Lodex, via TDM Factory (Gaillard et al., 2026), ou en ligne de commande (API)
- Prétraitement (*textExtract*), Classification (*aiAbstractCheck*), Validation (*bibCheck*), Extraction (*diseaseTag*), Alignement (*LoterreEnrich*)...

NB. Les précalculs prennent en compte l'ensemble du corpus ≠ les enrichissements s'appliquent document par document (ex. topRefExtract)

Fonctionnement
- Enrichissements



Interface administrateur, onglet ***Données, résultat de l'enrichissement***

Démonstration

Cas d’usage

Démonstration

Objectif

Construire et explorer un corpus sur les langues signées en SHS

Quelles sont les langues les plus étudiées ? Pourquoi ? Quelles sont les analyses effectuées (syntaxe, sémantique, sociolinguistique) ? Quels sont les auteurs ou travaux les plus cités ?

Démonstration
- Corpus

Istex Search

1 076 documents Istex : (title:("sign language" "sign languages" "signed language" "signed languages" "visuo-gestural language" "visuo-gestural languages" "visual-gestural language"...

Conditor

417 documents Conditor : ((title.en:("sign language" "signed language" "visuo-gestural language" "visual-gestural language"...

Démonstration - Déroulé

Importer deux jeux de données	Comment modifier un loader
Importer un modèle et configurer le thème	Comment modifier des éléments pré-existants
Créer un graphique	Comment créer un graphique à partir des mots-clés d'auteurs
Utiliser un enrichissement	Comment détecter les termes les plus spécifiques grâce à un enrichissement
Créer un enrichissement	Comment projeter une liste de termes issue de Glottolog
Créer un graphique en mode avancé	Comment configurer son graphique

Références

Cuxac, P. (2024, mai). La fouille de textes en IST : Les outils Istex-TDM. *INFORSID ' 24*. https://hal.science/hal-04597734

Eon, M., & Huguin, M. (2025). *Les publications de Laurent Schwartz en un corpus*. https://hal.science/hal-04991206

Gaillard, L., Bonvallot, V., Cuxac, P., & Parmentier, F. (2026). *TDM Factory : Rendre accessibles des algorithmes de fouilles de textes sans connaissances a priori ni paramétrages*. GC 2026, Jan 2026, Anglet, France. p.537-544.

Gregorio, S., Collignon, A., Parmentier, F., & Thouvenin, N. (2019, janvier). LODEX : Des données structurées au web sémantique. *Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*. https://hal.science/hal-01990444

Haut Conseil de la Santé Publique (2024). *Lutte contre les maltraitances des personnes en situation de vulnérabilité : Analyse et propositions du Haut Conseil de la santé publique* (Documents, p. 256). Haut Conseil de la Santé Publique.

Huguin, M., & Barreaux, S. (2023). *Le corpus « Machine Translation » : Une exploration diachronique des (méta)données Istex*. 18e Conférence en Recherche d'Information et Applications -- 16e Rencontres Jeunes Chercheurs en RI -- 30e Conférence sur le Traitement Automatique des Langues Naturelles -- 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. https://hal.science/hal-04131599

Huguin, M. (2025). *Comment créer, explorer et analyser un corpus de publications scientifiques avec Istex ?* 12èmes Journées de Linguistique de Corpus. https://cnrs.hal.science/hal-05332724