

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

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

Corpus Laurent Schwartz

Un autre regard sur les publications du mathématicien

Objectifs

Le corpus **Laurent Schwartz** permet aux chercheurs d'accéder à 271 travaux de Laurent Schwartz publiés ou réalisés entre 1936 et 2015. Ce corpus a fait l'objet de plusieurs enrichissements tant au niveau des textes que des métadonnées décrivant ces textes. On retrouve, par exemple, les codes de classement de la **classification MSC** ou encore des liens avec les **contenus du thésaurus de mathématiques de Loterie**. 15 filtres permettent de naviguer facilement dans le corpus. 14 graphiques dynamiques facilitent la visualisation des informations contenues dans les publications. Entre, 213 documents y sont directement accessibles en texte intégral.

En filigrane, les métadonnées des textes ainsi que les enrichissements apportés permettent de retracer la carrière scientifique du mathématicien. Après ses études de mathématiques à l'ENS, Laurent Schwartz s'engage dans une thèse intitulée *Étude des sommes d'exponentielles réelles* que l'on retrouve dans le corpus. Membre du groupe de mathématiciens Bourbaki, il est le premier français à obtenir la médaille Fields en 1950 pour ses travaux sur la théorie des distributions (cf. 1950 - 1951). Il enseigne successivement à l'université de Grenoble, de Nancy puis rejoint Paris, d'abord à la Sorbonne et à Polytechnique où il fonde le centre de recherche en mathématiques, aujourd'hui nommé **le centre de mathématiques Laurent Schwartz**. Retrouvez les différentes affiliations du chercheur dans le [graphique dédié](#).

(Eon & Huguin, 2025)

ISTEX Corpus

Des corpus scientifiques issus d'Istex pour inspirer vos projets de recherche

Corpus scientifiques / Corpus actualité / Collection Multidisciplinaire

Les publications Istex de la délégation Centre-Est du CNRS

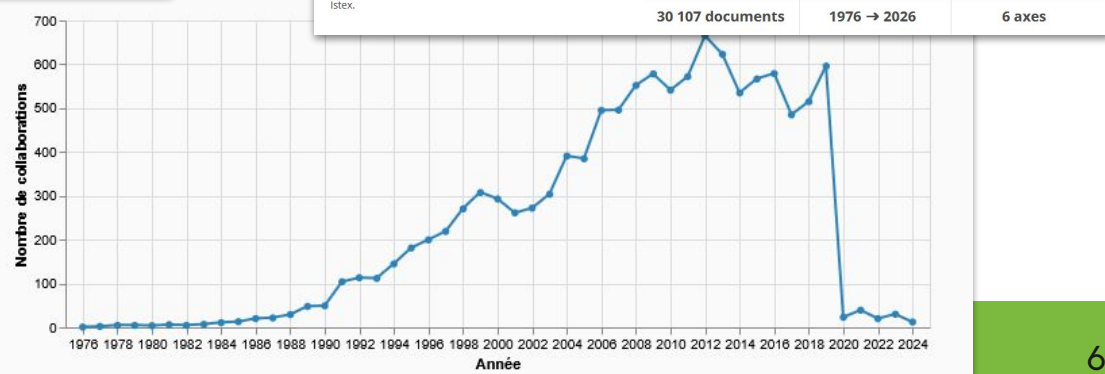
Plongez dans 50 ans de science en Centre-Est

Pour célébrer le cinquantenaire de sa Délégation Centre-Est, le CNRS et le CAES ont souhaité aller au-delà du simple hommage : retracer, données à l'appui, une histoire vivante de la recherche régionale. L'équipe [Istex](#) a ainsi constitué un premier corpus de publications analysées grâce aux [techniques de fouille de textes](#) de l'infrastructure Istex, opérée par [Pipist](#), institut du CNRS situé à Vandœuvre-lès-Nancy.



© CNRS Centre-Est - Service Communication

Le résultat ? Des graphiques interactifs qui donnent à voir, sur le temps long, les thématiques structurantes, les figures scientifiques marquantes et les grandes évolutions de la production en Centre-Est à travers le prisme du réservoir Istex.





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.

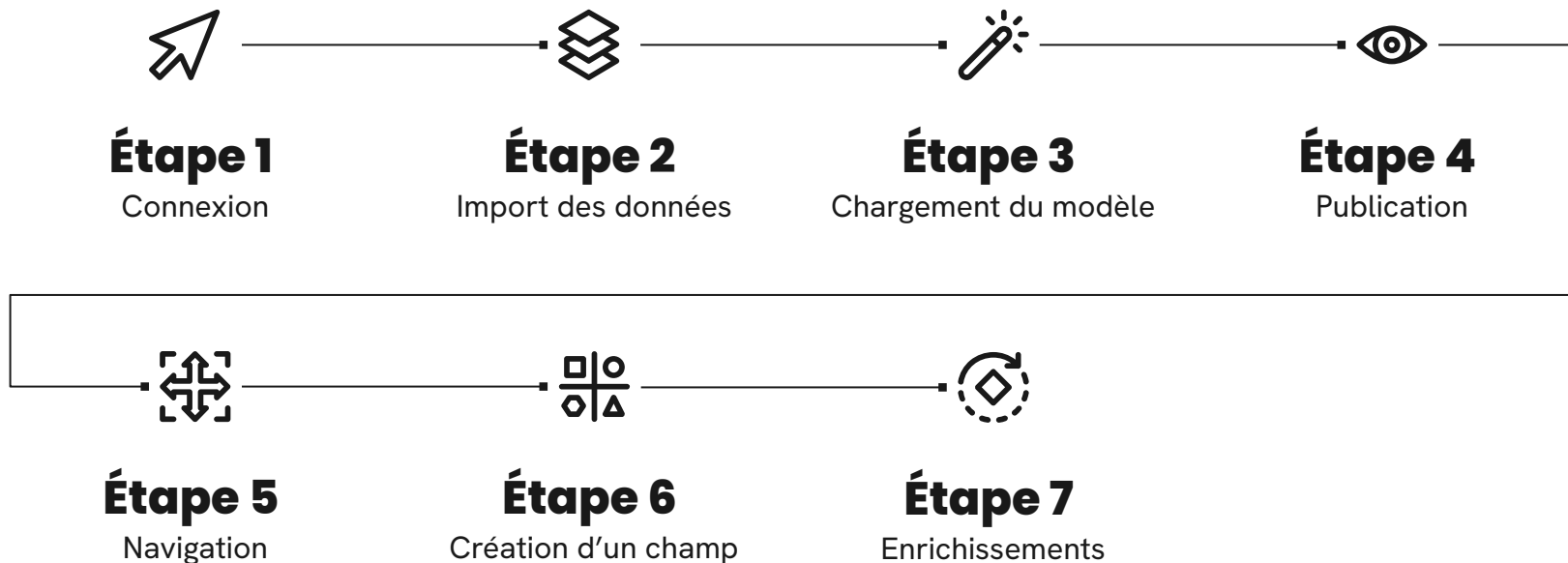


Site web

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.)

LODEX

Se connecter
pour modifier le contenu et l'apparence de ce site.

[GESTION DES INSTANCES](#)

Nom d'utilisateur *

Mot de passe *

CONNEXION

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



3 onglets principaux

LODEX

DONNÉES AFFICHAGE ANNOTATIONS PUBLIER

Données Enrichissements Précalculs

FICHIER LIEN WEB

Ajouter ou compléter votre instance en téléversant un fichier présent sur votre ordinateur. La taille

Glisser un fichier sur cette page ou cliquer sur l'icône de téléchargement

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

demo

- Modèle >
- Annotations >
- Avancé >
- Configuration
- Déconnexion

Choisir un loader Customiser le loader

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

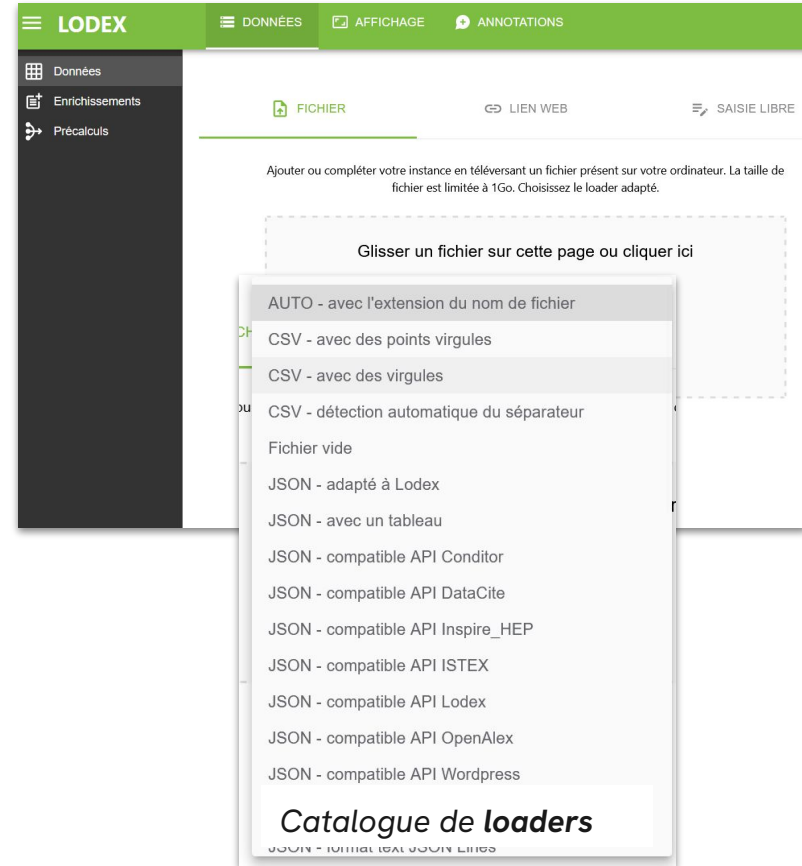


Fonctionnement

- 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



LODEX | DONNÉES | AFFICHAGE | ANNOTATIONS | PUBLIER

Données | Enrichissements | Précalculs

JEU DE DONNÉES | DONNÉES PRÉCALCULÉES

COLONNES | FILTRES | DENSITÉ | AJOUTER

<input type="checkbox"/>	uri	Titre	Auteur(s)	Affiliation(s)	Revue ou monogra...	Auteur(s) monogr
<input type="checkbox"/>	"ark:/67375/6H6-M4K1"	"Formalization of the s	["Jean-Baptiste Lamy"]	["LIMICS, Université F	"Knowledge-Based Sy	[]
<input type="checkbox"/>	"ark:/67375/6H6-MKH"	"Iconic languages: Tow	["Rita Francese","Mict	["Department of Comj	"Journal of Visual Lan	[]
<input type="checkbox"/>	"ark:/67375/6H6-ZNQI"	"How grammar can co	["Carlo Geraci","Marta	["Dipartimento di Psic	"Cognition"	[]
<input type="checkbox"/>	"ark:/67375/6H6-MLRl"	"Visual and linguistic c	["Evie Malaia","Ronnie	["Freiburg Institute for	"Cortex"	[]
<input type="checkbox"/>	"ark:/67375/6H6-FVCX"	"Analysis of the visual	["Rain G. Bosworth"],"C	["Department of Liber	"Vision Research"	[]
<input type="checkbox"/>	"ark:/67375/6H6-MXFf"	"An empirical investig	["Inge Zwitserlood"],"P:	["Radboud University	"Lingua"	[]
<input type="checkbox"/>	"ark:/67375/WNG-JK0"	"From Gesture to Sign	["Chloë R. Marshall"],"C	["Institute of Educatio	"Topics in Cognitive Si	[]

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

1-25 of 1367 < >

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



LODEX

DONNÉES AFFICHAGE ANNOTATIONS PUBLIER

Page d'accueil
Ressource principale
Sous-ressources
Graphiques
Recherche et facettes

PAGE DONNÉES PUBLIÉES

DATASET - Titre DATASET - Description + NOUVEAU CHAMP

Configurez votre affichage en cliquant sur le bouton « Nouveau champ »

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

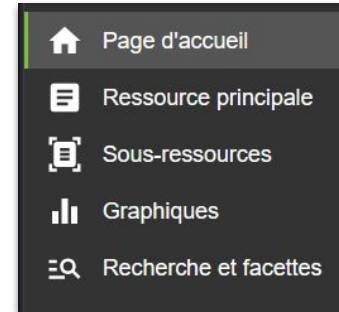


Fonctionnement

- 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

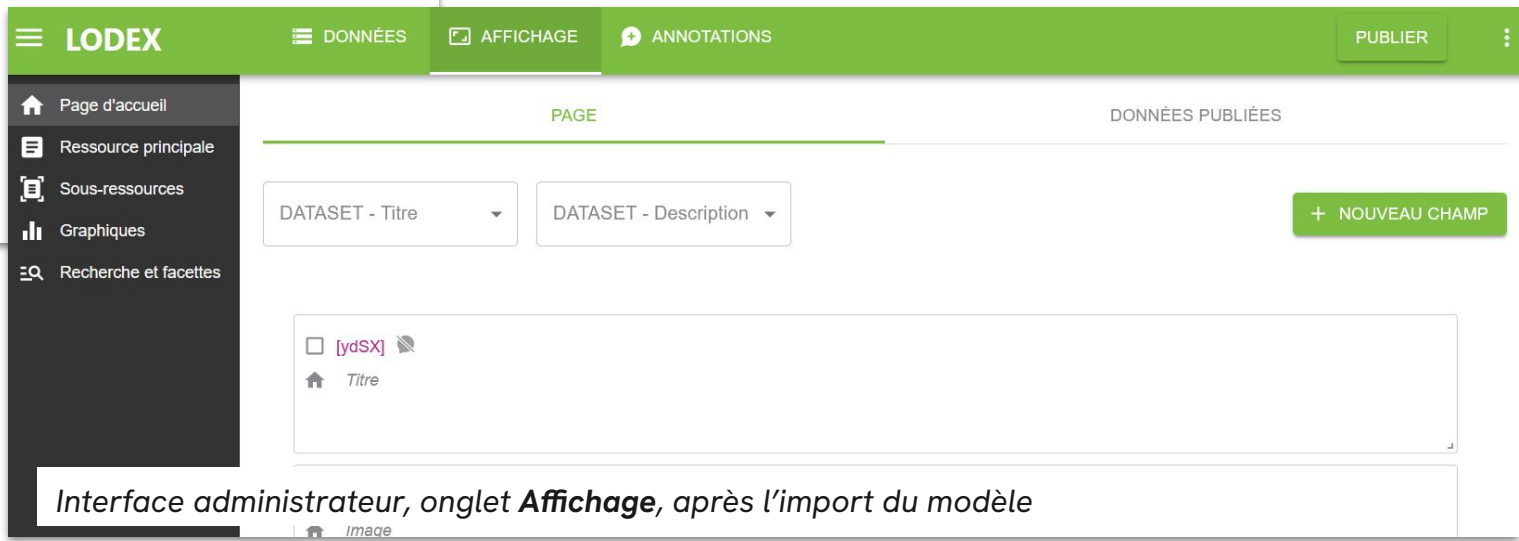




Le modèle est un fichier json ou tar.

Le site peut être publié.

le bouton « Nouveau champ »



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



Fonctionnement - Publication

The image displays two screenshots of the LODEX system. The left screenshot shows the administrative interface with a green header containing 'LODEX', 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS'. A sidebar on the left lists navigation options: 'Page d'accueil', 'Ressource principale', 'Sous-ressources', 'Graphiques', and 'Recherche et facettes'. The main content area is titled 'PAGE' and features two dropdown menus for 'DATASET - Titre' and 'DATASET - Description'. Below these is a card for a dataset with the identifier '[ydSX]' and a 'Titre' field. The right screenshot shows the 'Site publié' (published site) with a green header for 'ISTEX Démonstration'. The main content area features a section titled 'Langues signées' (Signed Languages) with the text 'La diversité des langues signées dans la recherche scientifique' and an illustration of five hands in different signing positions. Below this are three cards: 'Naviguer dans le site', 'Modifier le site', and 'Accéder aux données'. The footer of the published site includes 'Accueil', 'Graphiques', 'Recherche', and 'Voir Plus'.

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

- 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)



Étiquette
Mots-clés d'auteur(s)

Icône(s) du champ
Nom interne

Source de la valeur

VALEUR ARBITRAIRE CHOIX DE LA ROUTINE DONNÉE PRÉCALCULÉE COLONNE(S) EXISTANTE(S) DEPUIS UNE SOUS-RESSOURCE

Colonne(s) existante(s)
Mots-clés d'auteur Saisir colonne(s) existante(s)

Opérations de transformation TOUT SUPPRIMER

Aperçu de la valeur*

Mots-clés d'auteur(s)

["Ontology visualisation", "Querying task", "User centred design..."]

Visible

Afficher avec un format

Texte - Liste de valeurs

largeur 50 %

Légender un autre champ

Afficher en tant que champ composé

Champs composant l'affichage

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

LODEX

DONNÉES AFFICHAGE ANNOTATIONS

JEU DE DONNÉES

COLONNES FILTRES DENSITÉ AJOUTER

<input type="checkbox"/>	uri	Titre	Auteur(s)
<input type="checkbox"/>	"ark:/67375/6H6-M4K"	"Formalization of the s	["Jean-Baptiste Lamy"]
<input type="checkbox"/>	"ark:/67375/6H6-MKH"	"Iconic languages: Tov	["Rita Francese", "Mich
<input type="checkbox"/>	"ark:/67375/6H6-ZNQ"	"How grammar can co	["Carlo Geraci", "Marta
<input type="checkbox"/>	"ark:/67375/6H6-MLR"	"Visual and linguistic c	["Evie Malaia", "Ronnie
<input type="checkbox"/>	"ark:/67375/6H6-FVC"	"Analysis of the visual	["Rain G. Bosworth", "C
<input type="checkbox"/>	"ark:/67375/6H6-MXF"	"An empirical investig	["Inge Zwitserlood", "P
<input type="checkbox"/>	"ark:/67375/WNG-JK0"	"From Gesture to Sign	["Chloë R. Marshall", "C

Interface administrateur, onglet **Données**



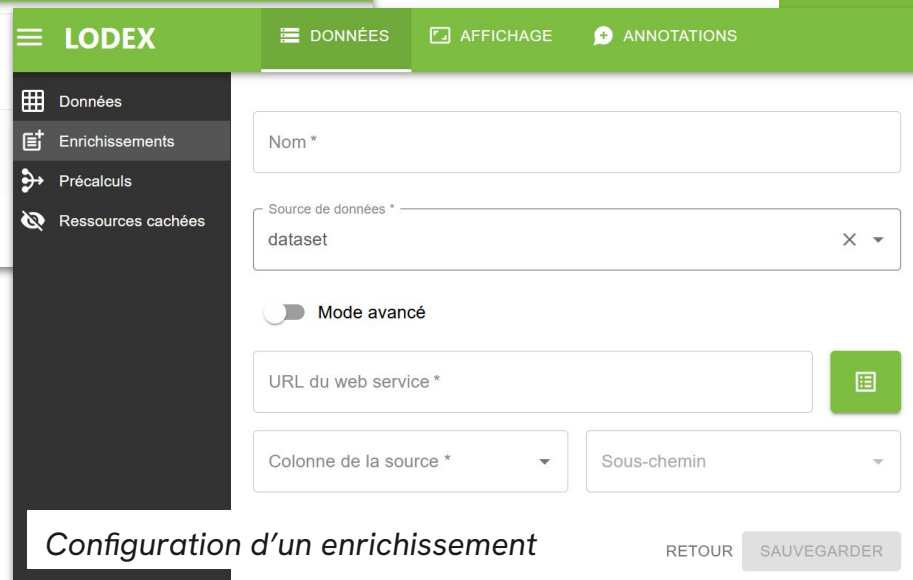
Fonctionnement

- Enrichissements



The screenshot shows the LODEX administrator interface. The top navigation bar is green and contains the LODEX logo, tabs for 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS', and a 'DÉPUBLIER' button. The left sidebar is dark grey with icons for 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main content area has a sub-header with 'COLONNES', 'FILTRES', 'DENSITÉ', 'AJOUTER', and 'LANCER TOUT'. Below this, there are columns for 'Nom', 'Colonne de la source', 'Sous-chemin', 'Mode avancé', and 'Statut'. The main area is currently empty, displaying 'Pas de résultats'.

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



The screenshot shows the LODEX configuration interface for an enrichment. The top navigation bar is green and contains the LODEX logo, tabs for 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS', and a 'DÉPUBLIER' button. The left sidebar is dark grey with icons for 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main content area is a form with the following fields: 'Nom *' (text input), 'Source de données *' (dropdown menu with 'dataset' selected), 'Mode avancé' (toggle switch), 'URL du web service *' (text input), 'Colonne de la source *' (dropdown menu), and 'Sous-chemin' (dropdown menu). There is a green 'Ajouter' button next to the 'URL du web service' field. At the bottom, there are 'RETOUR' and 'SAUVEGARDER' buttons.

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 catalogue** 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

The screenshot displays the LODEx administrator interface. The top navigation bar includes 'LODEX', 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS'. The left sidebar lists 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main content area shows the 'Données' tab with a search form for 'Extraction de termes spécifiques'. The search parameters include 'Nom*', 'Source de données*' (dataset), 'Mode avancé' (toggle), 'URL du web service*' (https://terms-extraction.services.istex.fr/v2/teeft/en), and 'Colonne de la source*' (Résumé). Below the search form, a table titled 'JEU DE DONNÉES' displays the results of the enrichment process. The table has columns for 'COLONNES', 'FILTRES', 'DENSITÉ', and 'AJOUTER'. The data rows show URIs and their corresponding enrichment results in JSON format.

COLONNES	FILTRES	DENSITÉ	AJOUTER
<input type="checkbox"/> uri			Extraction de termes spécifiques
<input type="checkbox"/> "ark:/67375/6H6-M4KW9CPT-Z"			[{"term": "iconic", "frequency": 5, "specificity": 1}, {"te
<input type="checkbox"/> "ark:/67375/6H6-MKHCWVHC-H"			[{"term": "iconic", "frequency": 3, "specificity": 1}, {"te

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





Démonstration

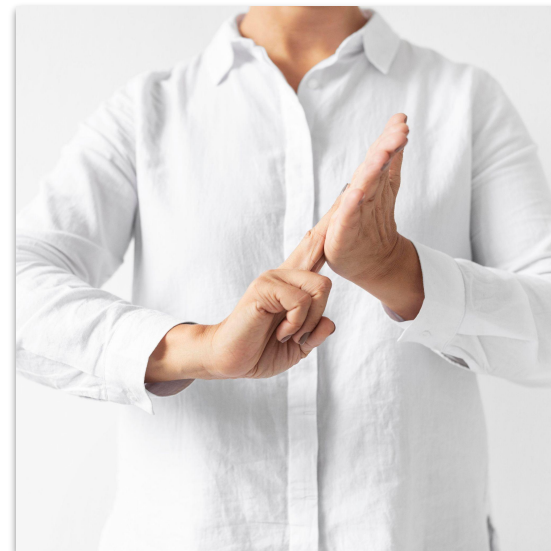
Cas d'usage

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

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

