# Corpus *Nom propre*

> ***"Je suis doctorant en linguistique, je débute une thèse sur le nom propre. Je souhaite repérer les revues de premiers plans, les thèmes récurrents, les notions clés, les auteurs et autrices incontournables sur le sujet et avoir une vision générale de l'étude du nom propre selon les disciplines."***

## 1. Description générale

Pour répondre à cette problématique, il faut constituer un corpus de documents traitant du nom propre. Istex, qui est un réservoir d'archives scientifiques mondiales, est une ressource particulièrement pertinente. Ce corpus va ensuite être exploré grâce à l'outil Lodex pour me fournir les informations les plus pertinentes (ex. les revues sur le sujet) et les ressources documentaires les plus judicieuses vis-à-vis de ma recherche.

- **Objectif** : repérer les études sur le nom propre.
- **Outil de TDM utilisé** : Lodex et les web services associés.
- **Contraintes imposées par l'outil** : Istex propose un format de sortie adaptée à l'outil Lodex. Pour utiliser les web services qui nous intéressent, il faut vérifier que les documents sont en anglais et contiennent des résumés.

## Exercice 1 – Construire une requête Istex

Étape 1 : Se rendre sur Istex Search : https://search.istex.fr.

Étape 2 : Rechercher les formes anglaises et françaises *nom propre*, *proper name* et *proper noun*.

- Pour une aide sur la syntaxe des requêtes consulter la documentation : https://doc.istex.fr/tdm/requetage/.
- Les espaces blancs sont considérés comme des opérateurs booléens OR (sauf en recherche assistée).
- Par défaut, la requête est insensible à la casse.
- Les guillemets permettent des recherches exactes (notamment avec des espaces). Ils sont inutiles en recherche avancée.

Étape 3 : Limiter le bruit et le silence.

- Rechercher les variantes morphosyntaxiques (singulier/pluriel) de chacun des termes de la requête.

- Par défaut, la recherche s'effectue dans tous les champs interrogeables d'Istex. Pour affiner la recherche, on peut préciser les champs les plus à même de renvoyer des résultats pertinents (soit le titre, le résumé et les mots-clés d'auteur dont la dénomination est *title*, *abstract* et *subject.value*). La requête prend alors la forme *champ:()*. La liste des champs est accessible dans l'assistant à la construction de requête.
- Pour répondre aux contraintes imposées par l'outil, il faut s'assurer de la présence de résumés dans les documents et s'assurer qu'ils sont en anglais.

***Je ne suis pas intéressé par les dictionnaires de noms propres (il faut aussi penser aux errata, index et back matter qui sont souvent des listes sans contenu scientifique). Je souhaite donc exclure les documents avec les termes* "dictionnaire" "dictionary" "index" "errata" *et* "back matter" *dans le titre.***

↳ *Les exclusions s'effectuent avec l'opérateur NOT.*

***Je m'intéresse exclusivement à la linguistique moderne : "en introduisant une distinction entre les analyses diachronique et synchronique du langage, Ferdinand de Saussure a jeté les bases de la linguistique moderne".***

↳ *Ferdinand de Saussure est mort en 1913. Peut-on limiter les dates de publication du corpus à des documents publiés après 1913 ?*

Étape 4 : Pour aller plus loin… transformer l'équation en utilisant des expressions régulières.

Les expressions régulières, présentées entre barres obliques (/), vous permettent de raccourcir votre requête.

- les . signifient n'importe quel caractère ;
- les * signifient n'importe quel nombre fois (s'appliquent au caractère précédent) ;
- les ? signifient entre 0 et 1 fois (s'appliquent au caractère précédent) ;
- les crochets [] sont l'équivalent d'un OR ;
- les .raw considèrent que le token interrogé est le champ dans sa globalité.

Pour en savoir plus et tester des expressions régulières se rendre sur https://regex101.com/.

## Exercice 2 – Premier pas vers le TDM

Étape 1 : Télécharger le corpus.

- Extraire le corpus *Nom propre* en utilisant l'équation définie et en choisissant le format adapté pour un import dans Lodex.

Étape 2 : Importer le corpus dans Lodex.

- Se rendre sur votre instance Lodex : se connecter avec votre nom d'utilisateur et votre mot de passe.
- Aller dans l'interface administrateur en cliquant sur "Voir plus > Admin".
- Importer le corpus en glissant le fichier *.zip* que vous venez de télécharger **sans décompression préalable**.
- Choisir un loader (script d'adaptation du fichier à Lodex qui dépend du format de fichier fourni en entrée). Choisissez le loader "ZIP résultat de dl.istex.fr". Cliquer sur "Importer les données".
- Un modèle est un fichier *.tar* qui permet de mettre en forme le site créé avec Lodex. Importer le modèle fourni , cliquer sur le menu en haut à droite "Modèle > Importer un modèle".
- Publier votre site pour une première visualisation en cliquant sur "Publier" en haut à droite. Cliquer sur l'icône en forme d'œil pour voir le résultat. Explorer les différents graphiques à partir de l'onglet "Graphiques" en bas. Consulter quelques ressources grâce à l'onglet "Recherche"pour vérifier que tout fonctionne.

Étape 3 : Extraction de mots-clés des résumés via le web service Teeft.

- Aller dans "Données > Enrichissements > + Ajouter", puis donner le nom "Mots-clés (WS)", aller chercher le web service **Teeft** approprié dans le catalogue (bouton vert à droite du champ "URL du web service"). Choisir "Résumé" dans la colonne de la source, cliquer sur "Sauvegarder". Cliquer enfin sur "Lancer".
- Dans Lodex, avant de faire un graphique, il est nécessaire de déclarer la colonne comme une ressource : dans "Affichage > Ressource principale", créer une ressource "Mots-clés Teeft" à partir de l'enrichissement "Mots-clés (WS)". On notera que l'**identifiant de la ressource** est une suite alphanumérique (sensible à la casse) de 4 caractères. L'identifiant est inscrit entre crochets à côté du nom de la ressource.
- Aller dans "Affichage > graphiques > + Nouveau champ", puis créer le graphique "Mots-clés les plus représentés dans les résumés" à partir de la ressource "Mots-clés Teeft". Choisir la routine "distinct-by" (dans général) puis ajouter derrière l'URL **l'identifiant de la ressource** "Mots-clés Teeft". Enfin, dans "Affichage", choisir le format "Diagramme en barres" en filtrant les résultats (mettre "valeur minimum à afficher" à 3).

Étape 4 : Utiliser le web service de TDM classification en domaines scientifiques.

- Aller dans “Données > Enrichissements > + Ajouter”, puis donner le nom “Catégories Inist (WS)”, aller chercher le web service approprié dans le catalogue (bouton vert à droite du champ “URL du web service”). Choisir la colonne la plus appropriée dans la colonne de la source, cliquer sur “Sauvegarder”. Cliquer enfin sur “Lancer”.
- L’enrichissement effectué apporte beaucoup d’informations. Pour ne garder que les domaines scientifiques, il faut nettoyer la colonne. Aller dans “Données > Enrichissements > + Ajouter”, puis donner le nom “Tri des catégories Inist (WS)”, cliquer sur “Mode avancé” et copier le code ci-dessous, cliquer sur “Sauvegarder”. Cliquer enfin sur “Lancer”.

```
[assign]
path = value
value = get("value.Catégories Inist (WS)").map((item)
=>`${item.rang} - ${item.code.value}`)
[assign]
path = value
value = fix('', self.value).filter(Boolean)
```

*Remarque : Le mode avancé offre une meilleure flexibilité pour transformer les données. Il n’est néanmoins pas nécessaire de comprendre le code. Les transformations les plus usuelles sont disponibles via ce lien.*

- Après avoir créé la ressource “Catégorie Inist” issue de l’enrichissement avancé “Tri des catégories Inist (WS)” (s’aider de l’étape 3), aller dans “Affichage > graphiques > + Nouveau champ”, puis créer le graphique “Domaines scientifiques” à partir de la ressource “Catégorie Inist” créée lors de l’étape 4. Choisir la routine “tree-by” (dans général) puis ajouter derrière l’URL l’identifiant de la ressource “Catégorie Inist”. Enfin, dans “Affichage”, choisir le format “graphique hiérarchique”.

Etape 5 : Parité

Dans cette section, nous allons faire une brève analyse de la parité dans le corpus en utilisant le web service de détection de genre sur les prénoms d’auteurs et d’autrices. En s’aidant du site Objectif TDM et des questions précédentes :

- Obtenir le nom du premier auteur ou de la première autrice en utilisant cet enrichissement avancé (en choisissant un nom de colonne approprié) :

```
[assign]
path = value
value = get("value.Auteur(s)").get(0)
```

- Lancer le web service de détection de genre sur la dernière colonne créée.

- Créer la ressource associée et un graphique permettant de représenter les genres retournés. On pourra s'aider de la documentation Lodex pour le choix de la routine.

Pour aller plus loin : croiser la parité avec la discipline scientifique.

- Faire tourner le web service de classification Hal sur le résumé. Créer la ressource associée.
- Utiliser cette ressource et celle du genre pour construire une carte de chaleur permettant de croiser la discipline et le genre.
  Indications :
  - Appliquer la transformation "GET" sur le path "labelFr" au moment de déclarer la ressource associée à la classification Hal.
  - Utiliser la routine "crossby" suivie des deux identifiants de ressource.