Atelier
Constitution et analyse de corpus

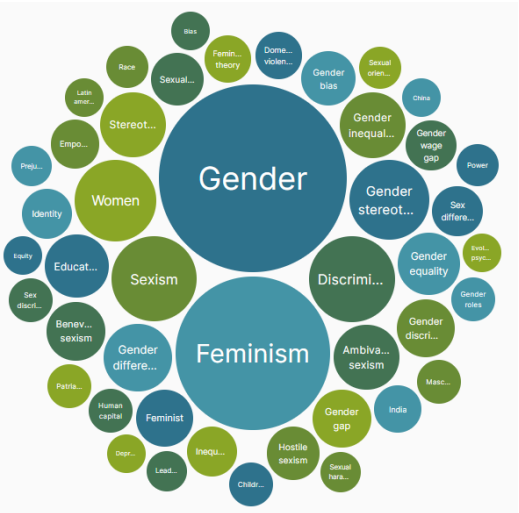
Comment la question des inégalités de genre est-elle abordée, représentée et évolue-t-elle dans la littérature scientifique selon les périodes et les zones géographiques ?

Description générale

Objectif : Pour répondre à cette problématique, il faut constituer un corpus de documents traitant de l'inégalité des genres en SHS grâce au réservoir Istex. Ce corpus va ensuite être exploré et enrichi dans Lodex.

Outils de TDM utilisés : Lodex et les web services associés.

Contraintes imposées par les outils : Istex Search propose un format de sortie adapté à Lodex. Pour utiliser les web services qui nous intéressent, il faut vérifier que les documents sont en anglais et contiennent des résumés.



Partie 1 – Requêtage

☐ **Étape 1 :** Se rendre sur Istex Search.

☐ **Étape 2 :** Rechercher les formes : *sexism, patriarchy, misogyny, feminism, gender discrimination, gender inequality, gender bias, gender gap, gender parity, gender equity, gender equality, gender stereotypes, sex discrimination, sexual discrimination*.

Au besoin consulter les *Astuces de recherche* : ? (à droite dans Istex Search).

☐ **Étape 3 :** Limiter le bruit et le silence.

⇒ Pour supprimer le silence : rechercher les variantes (singulier / pluriel, féminin / masculin et mots dérivés) de chacun des termes de la requête (ex. `feminis*` permet de rechercher les formes *feminist* et *feminism*).

⇒ Pour supprimer le bruit, on peut interroger le(s) champ(s) les plus à même de renvoyer des résultats pertinents (ex. le *titre*, le *résumé* et les *mots-clés d'auteur·ices* dont les dénominations sont *title*, *abstract* et *subject.value*). La requête prend alors la forme `champ:()`. La liste des champs est accessible dans la recherche assistée.

⇒ Pour limiter le bruit, on doit s'assurer de ne sélectionner que des articles de recherche (filtre : *Type de contenu*).

☐ **Étape 4 :** Répondre aux contraintes scientifiques.

⇒ Pour répondre à la problématique, il faut s'assurer de sélectionner des publications appartenant aux SHS.

☐ **Étape 5 :** Répondre aux contraintes techniques.

⇒ Pour répondre aux contraintes imposées par les outils, il faut s'assurer de la présence de résumés dans les documents et s'assurer qu'ils sont en anglais.

Partie 2 – Chargement des données et création du site Lodex

☐ **Étape 1 :** Télécharger le corpus.

⇒ Extraire le corpus en utilisant **l'équation corrigée** et en choisissant le format adapté pour un import dans Lodex.

☐ **Étape 2 :** Importer le corpus dans Lodex.

⇒ Se rendre sur **votre site** 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* téléchargé **sans décompression préalable**.
⇒ Choisir le loader[1] *ZIP Istex Search.*
⇒ Cliquer sur *Importer les données*.

☐ **Étape 3 :** Importer le modèle dans Lodex pour configurer un affichage minimal des données.

⇒ Importer le modèle fourni, cliquer sur le menu en haut à droite *Modèle > Importer un modèle*.
⇒ Configurer l'affichage : cliquer sur le menu en haut à droite *Configuration*. Choisir le thème *ISTEX – Thème ISTEX (restreint).* Cliquer sur *Sauvegarder*.
⇒ Publier votre site en cliquant sur *Publier* en haut à droite.
⇒ Cliquer sur l'icône en forme d'œil pour voir le résultat[2].
⇒ Explorer les différents graphiques à partir de l'onglet *Graphiques* en bas à gauche. Consulter quelques ressources grâce à l'onglet *Recherche*.

☐ **Étape 4 :** Créer un nouveau champ dans Lodex pour afficher les années de publication.

⇒ Depuis l'administration de l'instance se rendre sur l'onglet *Affichage*. Dans le menu de gauche, cliquer sur *Ressource principale*.
⇒ Cliquer sur le bouton *Nouveau champ*.
⇒ Dans le champ étiquette saisir : *Année*.
⇒ Dans la section *Source de la valeur* cliquer sur *Colonne(s) existante(s)*. Sélectionner la colonne *Date de publication*.
⇒ Cliquer sur le bouton *Sauvegarder*.

[1] Un loader est un script d'adaptation du fichier à Lodex. Il dépend du format de fichier fourni en entrée.
[2] Le texte présent dans le bandeau ainsi que les liens sont modifiables depuis la configuration.

☐ **Étape 5 :** Créer une nouvelle facette *Année*.

⇒ Depuis l'administration de l'instance se rendre sur l'onglet *Affichage*. Dans le menu de gauche cliquer sur *Recherche et facette*.
⇒ Dans la section Facettes sélectionner *Année*.
⇒ Sur la page d'accueil, depuis l'onglet recherche, vérifier que la facette a bien été créée.

? Combien de documents ont été publiés en 2000 ?

☐ **Étape 6 :** Créer un histogramme pour visualiser les années de publication.

⇒ Depuis l'administration de l'instance se rendre sur l'onglet *Affichage*. Dans le menu de gauche cliquer sur *Graphiques*.
⇒ Cliquer sur le bouton *Nouveau champ*.
⇒ Dans le champ étiquette saisir : *Années de publication*.
⇒ Dans la section *Source de la valeur* cliquer sur *Choix de la routine*[3]. Sélectionner la routine *distinct-by* et dans *Champs de la routine* choisir *Année*.
⇒ Dans *Affichage* choisir le format *Graphique – Diagramme en barres*. Dans *Paramètres des Données* trier par *Label Ascendant* et choisir la direction *vertical* (en dessous du jeu de couleurs). Cocher *Afficher l'info-bulle* et *Afficher les valeurs*.
⇒ Cliquer sur le bouton *Confirmer* puis *Sauvegarder*.
⇒ Vérifier que le graphique apparaît dans votre site.

[3] Une routine est un script ou un calcul permettant la création de graphique.

Partie 3 – Premier enrichissement grâce au DOI

Récupération des informations autour de l'open access via le web service Unpaywall[4].

☐ **Étape 1 :** Lancer le web service sur les DOI.

⇒ Aller dans *Données* > *Enrichissements* et cliquer sur *+ Ajouter*.

⇒ Donner le nom *doiEnrich_Unpaywall*, aller chercher l'url du web service lié à Unpaywall dans le catalogue en cliquant sur le bouton vert à droite du champ *URL du web service* (onglet *Métadonnées*).

⇒ Choisir *DOI* dans la *Colonne de la source*, cliquer sur *Sauvegarder*. À droite se trouve l'*Aperçu de la valeur*. Cliquer sur *Lancer*.

⇒ Dans *Données* > *Données*, une colonne enrichie (verte) est créée. Cliquer sur une cellule : les champs et les valeurs récupérés s'affichent, ils peuvent être modifiés au besoin.

☐ **Étape 2 :** Créer un diagramme circulaire pour visualiser les publications en open access.

Dans Lodex, avant de faire un graphique, il est nécessaire de déclarer la colonne comme une ressource.

⇒ Aller dans *Affichage* > *Ressource principale*, cliquer sur *+ Nouveau champ*.

⇒ Pour paramétrer ce nouveau champ : dans *Étiquette* nommer le champ ***OA_Unpaywall***, sélectionner *Colonne(s) existante(s)* et aller chercher la colonne précédemment créée et nommée *doiEnrich_Unpaywall*.

⇒ Le web service Unpaywall renvoie, pour chaque document, plusieurs champs dont ceux concernant l'accès libre. Dans notre cas, nous souhaitons récupérer la valeur du champ *is_oa* : cliquer sur *Ajouter une opération*, sélectionner **GET**, puis indiquer *is_oa* dans le champ *path*.

⇒ L'aperçu de la valeur montre des valeurs sans guillemets. Il s'agit de booléens non considérés comme chaînes de caractères. Pour que ces valeurs soient interprétables, une autre opération est nécessaire : sélectionner **STRING**. Les valeurs sont désormais entre guillemets.

⇒ Cliquer sur *Sauvegarder*.

⇒ Pour créer le graphique, aller dans *Affichage* > *graphiques* cliquer sur *+ Nouveau champ*.

⇒ Nommer le graphique en renseignant *Open Access (Unpaywall)* dans le champ *Étiquette*. Choisir la routine ***distinct-by*** puis choisir le champ sur lequel la routine va s'appliquer grâce au menu déroulant (soit le champ *OA_Unpaywall*).

[4] *Unpaywall* est une base qui recense des métadonnées de publications scientifiques.

⇒ Enfin, dans *Affichage*, choisir le format ***Graphique – Diagramme circulaire***.

⇒ Cliquer sur *Confirmer* puis sur *Sauvegarder*.

❓ Combien de documents sont en libre accès ? Quelle différence constatez-vous avec les chiffres Istex Search ?

Partie 4 – Détection du genre des auteurs

Détection du genre de l'auteur à partir du prénom à l'aide du web service genderDetect.

☐ **Étape 1 :** Lancer le web service *genderDetect*.

⇒ Reproduire la procédure pour créer un nouvel enrichissement. Donner le nom *genderDetect*, aller chercher l'url du web service **genderDetect** dans le catalogue (onglet *Classification*).

⇒ Choisir *Auteur(s)* dans la *Colonne de la source.*

⇒ Cliquer sur *Sauvegarder* puis *Lancer*. Attendre la fin du traitement [environ 5 minutes].

☐ **Étape 2 :** Créer une carte proportionnelle (*treemap*) pour visualiser les genres détectés.

⇒ Aller dans *Affichage* > *Ressource principale*, cliquer sur *+ Nouveau champ*.

⇒ Dans *Étiquette* nommer le champ *Genre des auteur·ices*, sélectionner *Colonne(s) existante(s)* et aller chercher la colonne nommée *genderDetect*. Paramétrer l'affichage du champ : dans *Affichage* choisir *Texte - Liste de valeurs* dans le catalogue. Cliquer sur *Confirmer*.

⇒ Cliquer sur *Sauvegarder*.

⇒ Créer la facette associée.

⇒ Pour créer le graphique : aller dans *Affichage* > *Graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Genre des auteur·ices* dans le champ *Étiquette.*

⇒ Choisir la routine ***distinct-by*** puis choisir le champ sur lequel la routine va s'appliquer grâce au menu déroulant (soit le champ G*enre des auteur·ices*). Enfin, dans *Affichage*, choisir le format ***Graphique – Carte proportionnelle***.

⇒ *Dans Paramètres des Données* choisir le tri *Valeur Descendant* et mettre le *Nombre max de champs* à 10. Décocher *Données hiérarchiques.*

⇒ Cliquer sur *Confirmer* puis sur *Sauvegarder*.

? Les femmes publient-elles plus sur les discriminations liées au genre ?

☐ **Étape 3 :** Créer une carte de chaleur pour visualiser les thématiques abordées selon le genre.

⇒ Pour créer le graphique : aller dans *Affichage* > *Graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Thématiques abordées selon le genre des auteur·ices* dans le champ *Étiquette.*

⇒ Choisir la routine ***cross-by*** puis choisir les champs sur lesquels la routine va s'appliquer grâce au menu déroulant (soit les champs G*enre des auteur·ices* et *Mots-*

clés d'auteur·ices). Enfin, dans *Affichage*, choisir le format ***Graphique – Carte de chaleur***.

⇒ *Dans Paramètres des Données choisir le tri Valeur Descendant* et mettre le *Nombre max de champs* à 30.

⇒ Cliquer sur *Confirmer* puis sur *Sauvegarder*.

Partie 5 – Création d'une cartographie des pays publiants

Utilisation des affiliations pour projeter les pays de publication sur une carte.

☐ **Étape 1 :** Créer un enrichissement pour découper les affiliations.

⇒ Aller dans *Données* > *Enrichissements* et cliquer sur *+ Ajouter*.

⇒ Donner le nom *addressSplit*, aller chercher l'url du web service ***addressSplit*** dans le catalogue en cliquant sur le bouton vert à droite du champ *URL du web service*.

⇒ Choisir *Affiliation(s)* dans la *Colonne de la source* puis cliquer sur *Sauvegarder*.

⇒ Dans l'*Aperçu des données*, on visualise un tableau de tableaux `[[tableau1],[tableau2]]`.

⇒ Pour que les données soient correctement traitées, il est nécessaire d'avoir un tableau. Pour ce faire, actionner le *Mode avancé*, pour visualiser les lignes de commandes Lodash. Remplacer la ligne 11 par le contenu ci-dessous.

```
value = get("value.Affiliation(s)").flatten()
```

Remarque : le mode avancé offre une meilleure flexibilité pour transformer les données. Il s'agit de code en Lodash, une librairie Javascript (cf. tutoriel). Les transformations les plus usuelles sont disponibles via ce lien.

⇒ *Sauvegarder* et *Lancer*.

☐ **Étape 2 :** Créer une cartographie.

⇒ Dans *Affichage* > *Ressource principale*, créer une ressource *Pays des affiliations* à partir de l'enrichissement. Après avoir sélectionné la colonne, cliquer sur *Ajouter une opération* pour appliquer une transformation **GET**, renseigner *value.country* dans le *path*. Cliquer sur *Sauvegarder*.

⇒ Créer une facette associée.

⇒ Pour créer le graphique : aller dans *Affichage* > *graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Cartographie des pays publiants* dans le champ *Étiquette*.

⇒ Choisir la routine ***distinct-ISO3166-1-alpha3-from*** puis choisir le champ sur lequel la routine va s'appliquer.

⇒ Enfin, dans *Affichage*, choisir le format *Cartographie, paramétrer les données pour les trier en Descendant*. Cliquer sur *Confirmer* puis *Sauvegarder*.

❓ Quel pays publie le plus ? Quel problème constatez-vous en examinant les valeurs de la facette *Pays des affiliations* ?

Partie 6 – Uniformiser les informations géographiques

☐ **Étape 1 :** Créer un enrichissement pour récupérer le pays des affiliations.

⇒ Aller dans *Données* > *Enrichissements* et cliquer sur *+ Ajouter*.

⇒ Donner le nom *Pays des affiliations*, passer en Mode avancé et coller le code ci-dessous.

```
[assign]
path=value
value=get("value.addressSplit").map('value.country')
```

⇒ *Sauvegarder* et *Lancer*.

☐ **Étape 2 :** Lancer l'enrichissement Pays et subdivisions.

⇒ Aller dans *Données* > *Enrichissements* et cliquer sur *+ Ajouter*.

⇒ Donner le nom *Loterre*, aller chercher l'url du web service ***Pays et subdivisions*** dans le catalogue en cliquant sur le bouton vert à droite du champ *URL du web service*.

⇒ Choisir *Pays des affiliations* dans la *Colonne de la source* puis cliquer sur *Sauvegarder et Lancer.*

☐ **Étape 3 :** Créer une cartographie de flux des pays co-publiants.

⇒ Dans *Affichage* > *Ressource principale*, créer une ressource *Loterre* à partir de l'enrichissement. Après avoir sélectionné la colonne, cliquer sur *Ajouter une opération* pour appliquer une transformation **GET**, renseigner *cartographyCode* dans le *path*. Cliquer sur *Sauvegarder*.

⇒ Pour créer le graphique : aller dans *Affichage* > *graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Cartographie des pays co-publiants* dans le champ *Étiquette.*

⇒ Choisir la routine ***payring-with*** puis choisir les champs sur lequel la routine va s'appliquer : *Loterre* et *Loterre*.

⇒ Enfin, dans *Affichage*, choisir le format *Cartographie de flux, paramétrer les données pour les trier en Descendant*. Cliquer sur *Confirmer* puis *Sauvegarder*.

Partie 7 – Projection d'une terminologie

Utilisation du mode avancé pour détecter la présence de termes du thésaurus de l'European Institute for Gender Equality dans les mots-clés.

☐ **Étape 1 :** Regrouper les mots-clés.

⇒ Créer un nouvel enrichissement pour fusionner les mots-clés d'auteurs et les mots-clés Teeft : le nommer *Tous les mots-clés* passer en *Mode avancé* et coller le code ci-dessous.

```
[assign]
path = value
value = fix(self.value["Mots-clés d'auteur"], self.value["Mots-clés
(teeft)"]).join(',').toLower().replace(/ \(...\)/g,"").split(',')
```

⇒ *Sauvegarder* et *Lancer* cet enrichissement.

☐ **Étape 2 :** Projeter la terminologie sur la liste pour détecter la présence de concepts clés.

⇒ Créer un nouvel enrichissement : le nommer *Termes détectés* et coller le code ci-dessous. Cliquer sur *Sauvegarder* puis *Lancer*.

```
# Création d'une variable d'environnement qui stocke la date au format
ISO
[env]
path = date
value = thru(d => (new Date()).toISOString().split("T")[0])

# Choix du champ Lodex à rechercher dans le dictionnaire
[assign]
path = value
value = get("value.Tous les mots-clés")

# Sert à traiter individuellement chaque valeur de la liste (chaque code
ici)
[map]
path = value

# Création d'une clé temporaire pour stocker la valeur actuelle
[map/replace]
path = tmpkey
value = self()

[map/combine]
path = tmpkey

# URL vers un fichier TSV accessible via internet
primer = https://bibliotheque-supports-
1.formation.lodex.fr/terminologie-ANF.tsv
```

```
# Création d'un fichier temporaire local avec la date définie dans [env]
(facultatif, évite de télécharger le fichier plusieurs fois)
cacheName = env("date").prepend("Variable")
default=

[map/combine/URLStream]
path = false

[map/combine/CSVParse]
separator = fix("\t")

[map/combine/CSVObject]

[map/combine/replace]
path = id
# Choix de la colonne du fichier TSV à mettre en correspondance
value = get('Term').toLower()
path = value
value = self()

[map/exchange]
value = get('tmpkey.value')

[assign]
path = value
value = get("value").compact()
```

☐ **Étape 3 :** Créer un graphique de flux.

⇒ Dans *Affichage* > *Ressource principale*, créer une ressource *Termes détectés* à partir de l'enrichissement. Après avoir sélectionné la colonne, cliquer sur *Ajouter une opération* pour appliquer une transformation **GET**, renseigner *Term* dans le *path*. Cliquer sur *Sauvegarder*.

⇒ Pour créer le graphique : aller dans *Affichage* > *graphiques* cliquer sur *+ Nouveau champ*, nommer le graphique en renseignant *Évolution des concepts cités* dans le champ *Étiquette.*

⇒ Choisir la routine ***cross-by*** puis choisir les champs sur lequel la routine va s'appliquer : à savoir *Année* et *Termes détectés*.

⇒ Enfin, dans *Affichage*, choisir le format *Graphique de flux, paramétrer les données pour les trier en Descendant, passer le Nombre max de champs à 2000*. Cliquer sur *Confirmer* puis *Sauvegarder*.

? En quelle année apparaît le terme *heterosexism* dans le corpus ?

Partie 8 – Premier précalcul

Détecter les articles les plus cités dans le corpus.

☐ **Étape 1 :** Lancer le précalcul topRefExtract.

⇒ Aller dans *Données* > *Précalculs* et cliquer sur *+ Ajouter*.

⇒ Donner le nom *topRef*, aller chercher l'url du précalcul topRefExtract dans le catalogue en cliquant sur le bouton vert à droite du champ *URL du web service*.

⇒ Choisir DOI dans la *Colonne de la source* puis cliquer sur *Sauvegarder et Lancer.*

? Quels sont articles les plus cités ?