



Constitution de corpus - Défis méthodologiques et solutions apportées par l'infrastructure Istex

Mathilde Huguin

► To cite this version:

Mathilde Huguin. Constitution de corpus - Défis méthodologiques et solutions apportées par l'infrastructure Istex. École thématique. Modèles de langue pour le traitement sémantique et l'intégration de connaissances et données en agriculture, alimentation et environnement, Montpellier, France. 2025. hal-05332761

HAL Id: hal-05332761

<https://cnrs.hal.science/hal-05332761v1>

Submitted on 27 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Constitution de corpus

Défis méthodologiques et solutions apportées par l'infrastructure Istex

Mathilde Huguin

IR PhD Inist - CNRS

mathilde.huguin@inist.fr

2025 | École thématique *Modèles de langue pour le traitement sémantique et l'intégration de connaissances et données en agriculture, alimentation et environnement*



Inist- CNRS

Plan

PARTIE 1 : ÉLÉMENTS CONTEXTUELS

1. Les corpus
2. Défis méthodologiques
3. Istex

PARTIE 2 : CONSTITUTION D'UN CORPUS SPÉCIALISÉ

4. Présentation du cas d'usage
5. Sélection des documents
6. Exploration, visualisation & affinage
7. Premières annotations : extractions d'EN
8. Conclusion
9. Annexes et références

PARTIE 1 : ÉLÉMENTS CONTEXTUELS



Les corpus



Les corpus

Qu'est ce qu'un corpus ?

Partie 1 : éléments contextuels | Les corpus

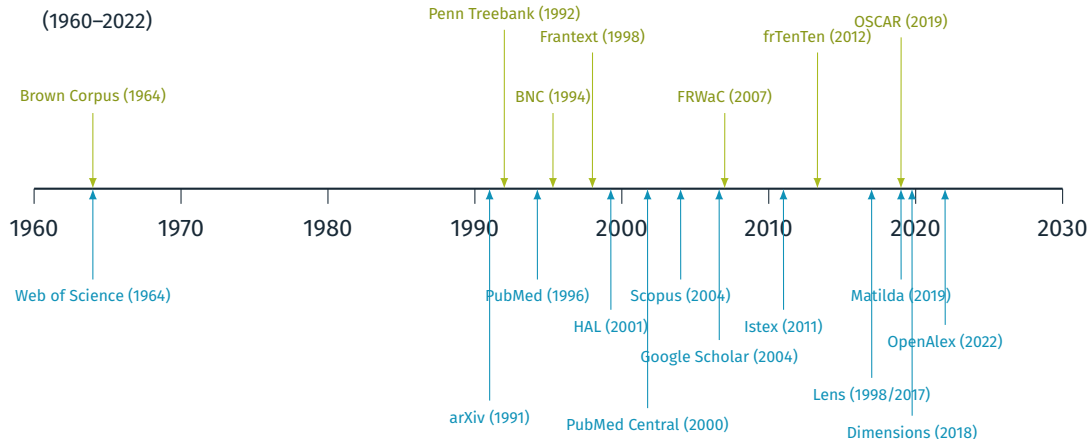
QU'EST-CE QU'UN CORPUS ?

- Employé depuis l'antiquité dans la langue savante pour désigner des collections de textes (juridiques, épigraphiques, ...) : *"Recueil formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline"* (Mellet, 2002)
- Apparition du terme chez les linguistes dans les années 60 alors que des collectes existent depuis longtemps (premiers corpus informatisés; cf. Léon, 2008; Benzitoun et Cappeau, 2025)
- Définition très discutée en linguistique et en **linguistique de corpus** (taille, représentativité, etc. cf. Sinclair, 1996; Habert, 2000); distinction entre réservoir (ou archive) et corpus (Rastier, 2004)
- Depuis 2000, évolution exponentielle des réservoirs et des corpus (documentaires ou non)
- Matière première des LLM : ils structurent, apportent savoir, diversité linguistique et contextuelle, conditionnent l'efficacité des modèles
- Utilisation des LLM sur les corpus (augmentation de 47% des publications en 6 ans, entre 12 et 82% des publications jamais lues, cf. Larivière *et al.*, 2009)

Partie 1 : éléments contextuels | Les corpus

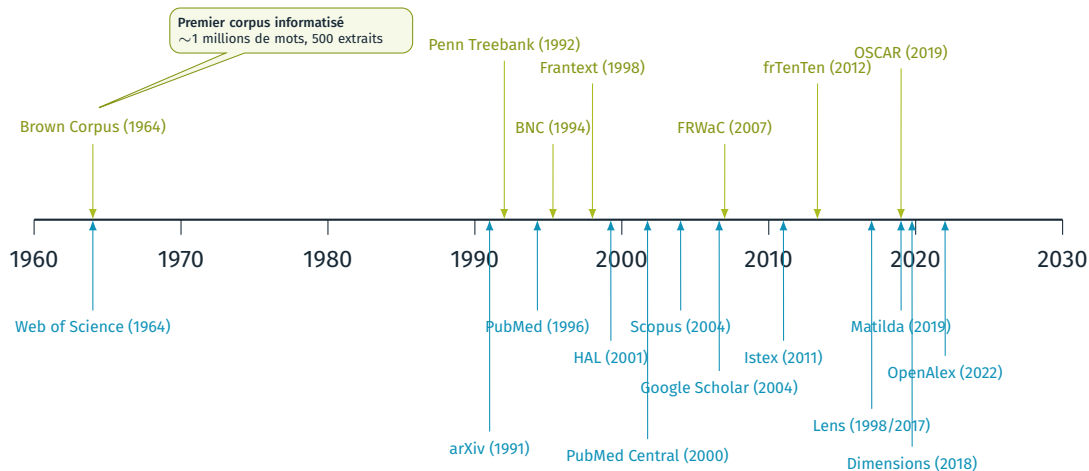
QU'EST-CE QU'UN CORPUS ?

Évolution des corpus et bases de données documentaires (1960–2022)



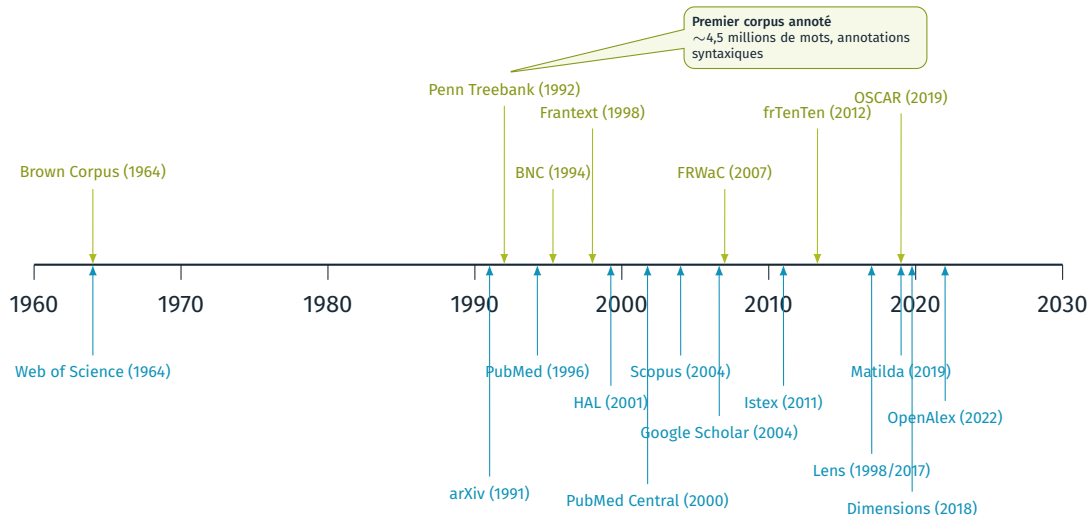
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



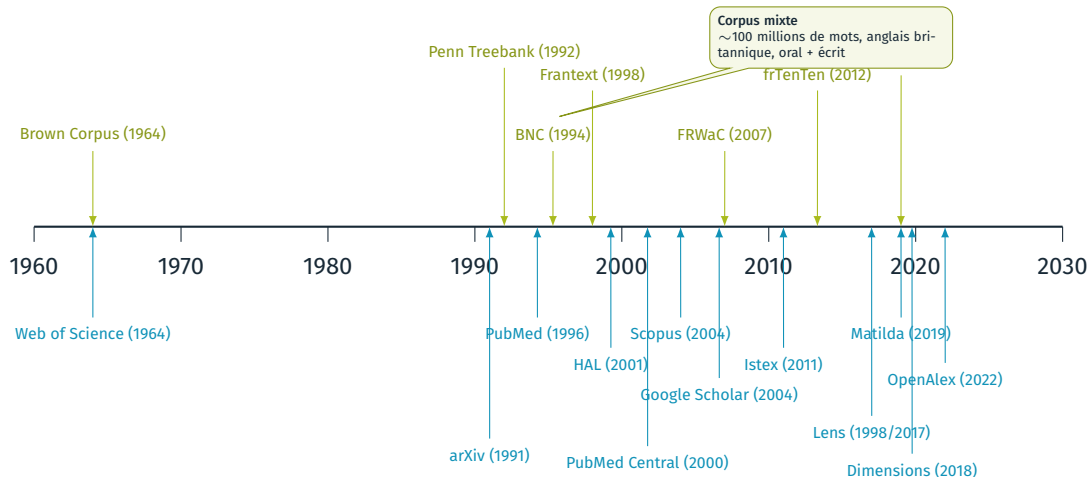
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



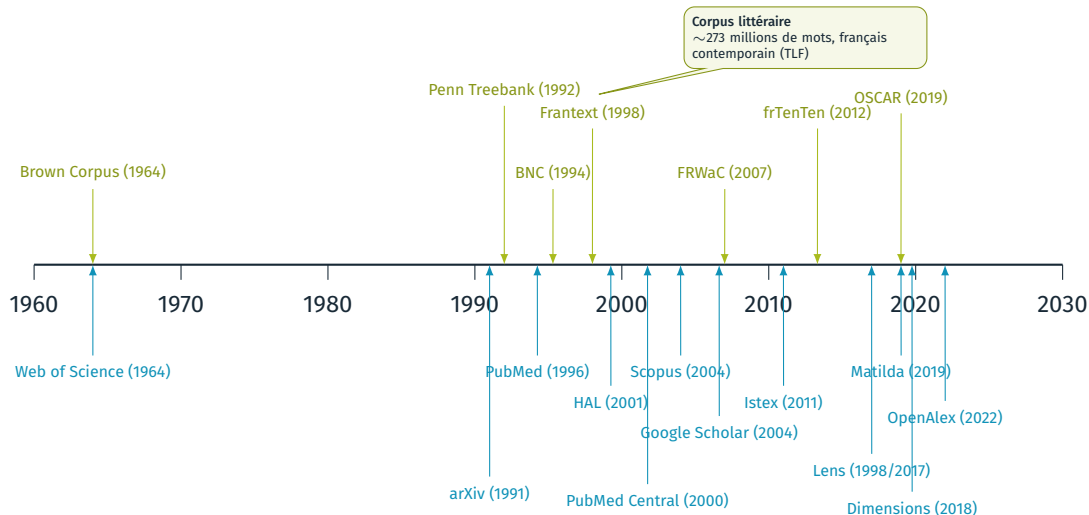
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



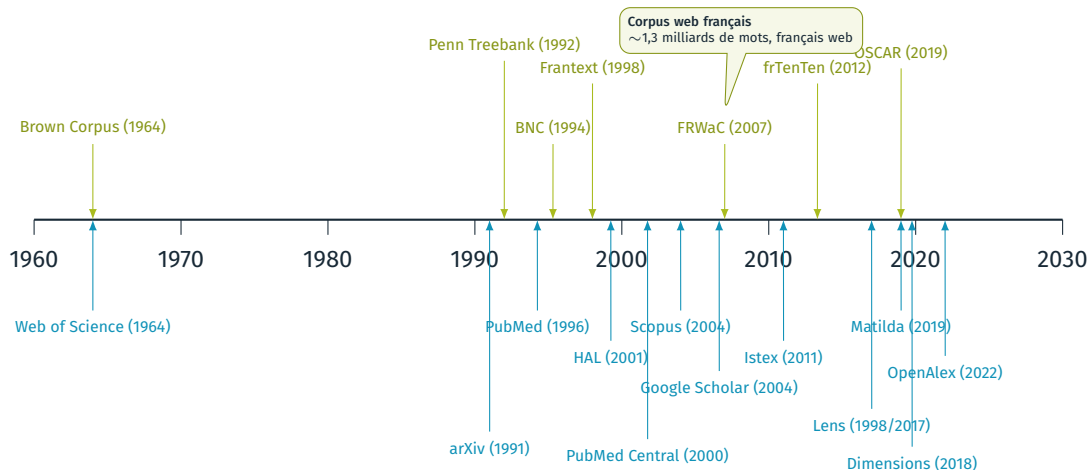
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



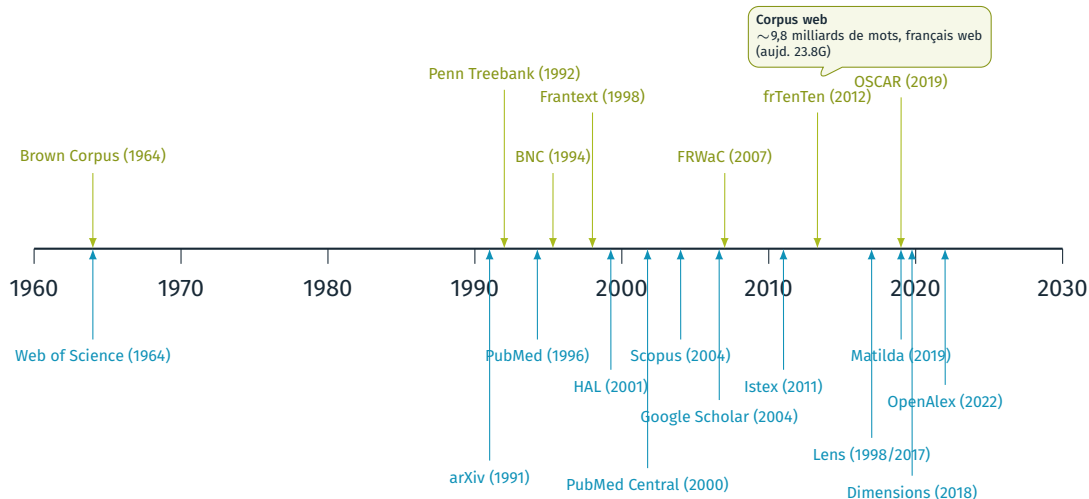
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



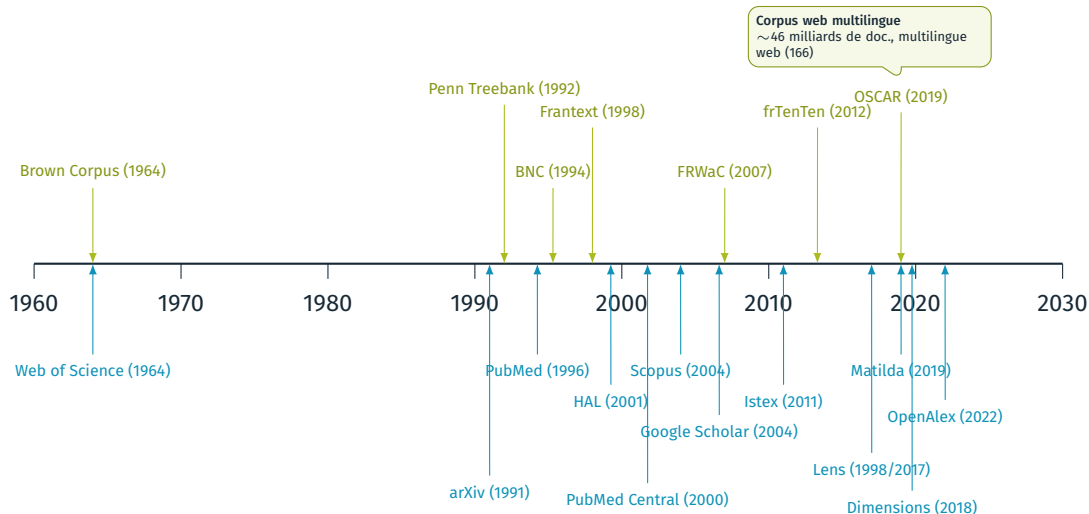
Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?



Partie 1 : éléments contextuels | Les corpus

QU'EST-CE QU'UN CORPUS ?





Les corpus

Corpus & LLM

Partie 1 : éléments contextuels | Les corpus

CORPUS & LLM

Type de corpus	Fonction	Exemple
Pre-training Corpora	Apprentissage initial du modèle sur un vaste corpus généraliste multi-domaines	Donner au modèle une connaissance générale du monde → Common Crawl pour ChatGPT (données massives issues du web)
Instruction Fine-tuning Datasets	Adaptation du modèle via des exemples supervisés ou consignes explicites	Répondre à des instructions claires → Alpaca (52 000 instructions-réponses générées par GPT-3.5, utilisé pour fine-tuner LLaMA)
Preference Data-sets	Apprentissage des préférences utilisateur, renforcement par feedback humain	Privilégier les réponses utiles → Stiennon <i>et al.</i> (2022) (réponse A plus pertinente que B)
Evaluation Data-sets	Évaluation de la qualité, précision et robustesse du modèle à l'aide de benchmarks standardisés	Tester la compréhension multi-domaines → MMLU (questions multi-choix couvrant 57 domaines, cf. Hendrycks <i>et al.</i> , 2021)

Usages majeurs selon Liu *et al.* (2024)



Les corpus

Biais

Partie 1 : éléments contextuels | Les corpus

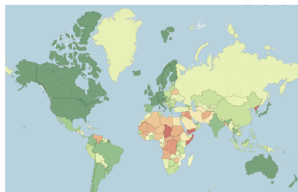
BIAIS

Exemples de biais induits par les datasets

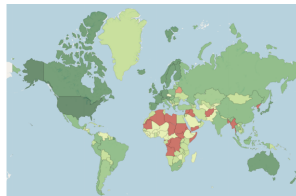
- IA du "pôle emploi" autrichien : *"Le Berufsinfomat a par exemple proposé à des candidats masculins de postuler dans le domaine de l'informatique, tandis qu'ils conseillaient à des candidates ayant un CV identique de se tourner vers les études de genre ou la restauration"* (cf. Courrier international 2024)
- Certains outils médicaux basés sur l'IA (OpenAI's GPT-4, Meta's Llama 3 and Palmyra-Med) ont tendance à minimiser les symptômes des femmes ou des minorités ethniques (cf. Financial Times 2025)

Partie 1 : éléments contextuels | Les corpus

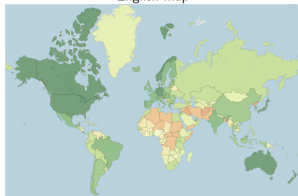
BIAIS



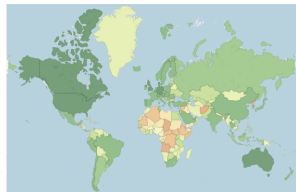
English map



French map



Russian map



Arabic map

« Mapping the answers to the prompt "Could you assign a positivity score to these words" for every country in the world, in 4 languages » (Boussidan *et al.*, 2024)



Défis méthodologiques



Défis méthodologiques

Questions incontournables

Partie 1 : éléments contextuels | Défis méthodologiques

QUESTIONS INCONTOURNABLES

La constitution d'un corpus peut être plus longue que son analyse

- Quels **genres de texte** doit-on inclure dans le corpus (scientifique, journalistique, conversationnel, littéraire, etc.)? Comment gérer la **qualité linguistique** des données (erreurs, bruit, doublons, textes tronqués, OCR de mauvaise qualité)?
- Quels sont les **droits et licences associés** aux données, et peut-on les utiliser légalement pour l'entraînement ou la recherche?
- Sous quels **formats** les données sont-elles disponibles (texte brut, PDF, HTML, XML, JSON, etc.), et sont-ils adaptés au traitement automatique?
- Quelle **méthode de récupération des documents** est employée (web scraping, API, dépôts institutionnels, partenariats, crowdsourcing), et quelles en sont les limites?

Partie 1 : éléments contextuels | Défis méthodologiques

QUESTIONS INCONTOURNABLES

Genre	Atouts	Risques et limites
Juridique	Précision, structure, faible ambiguïté	Faible variété, manque d'expression naturelle et d'exemples pratiques
Web	Grande diversité thématique, registres variés	Qualité inégale, bruit, duplication, informations erronées
Journalistique	Clarté, structuration éditoriale, capacité argumentative	Biais idéologique/politique, diversité de formats, subjectivité
Scientifique	Rigueur factuelle, données structurées, terminologie experte	Jargon/terminologie, spécialisation, généralisation limitée
Brevet	Descriptions techniques détaillées, structure normée	Complexité technique, difficulté d'interprétation

Influence du genre de texte sur la qualité du corpus (e.g. Labadie et Prince (2008) sur la segmentation)

Partie 1 : éléments contextuels | Défis méthodologiques

QUESTIONS INCONTOURNABLES

Droits et licences associés aux données

- Textes littéraires : copyright souvent strict (ex. éditeurs modernes), Open Access rare (Project Gutenberg pour œuvres tombées dans le domaine public)
- Données du web : protégées par copyright, conditions d'utilisation souvent restrictives (sauf API ouvertes)
- Journaux quotidiens : protégés par copyright, reproduction interdite
- Publications scientifiques : copyright par l'éditeur, TDM possible pour la recherche (**directive UE 2019**; transposition de la directive européenne en droit français 2020/2021)

Partie 1 : éléments contextuels | Défis méthodologiques

QUESTIONS INCONTOURNABLES

Formats

CSV adapté à des contenus courts ou tabulaires mais ne gère pas la hiérarchie du texte et peut être limitant pour des documents complexes

PDF, DOCX extractions sont souvent imparfaites : risques liés au formatage, aux éléments non textuels et à la segmentation non standardisée

TXT universel et simple à manipuler mais sans structure balisée, perte des tableaux ou figures

XML/TEI facilite l'extraction structurée mais nécessite souvent des parsers complexes, et la gestion de schémas différents selon les sources peut s'avérer fastidieuse

JSON permet de structurer les textes et les métadonnées mais requiert un nettoyage intensif si la structure varie

Partie 1 : éléments contextuels | Défis méthodologiques

QUESTIONS INCONTOURNABLES

Moyen	Description / Avantages	Limites	Exemple
API	Accès structuré aux données, rapide et fiable	Limité en volume et quotas	HAL, arXiv API, Twitter API
Web scraping	Extraction flexible depuis des pages web	robots.txt bloquant, nettoyage nécessaire	Articles de presse, blogs, forums
Interface web	Téléchargement direct de lots de documents	Conversion parfois nécessaire (PDF → texte), métadonnées limitées	Project Gutenberg, PubMed Central
Crowdsourcing	Création de corpus via contributions humaines	Temps et coût humain, contrôle de qualité	

Moyens techniques pour récupérer des données

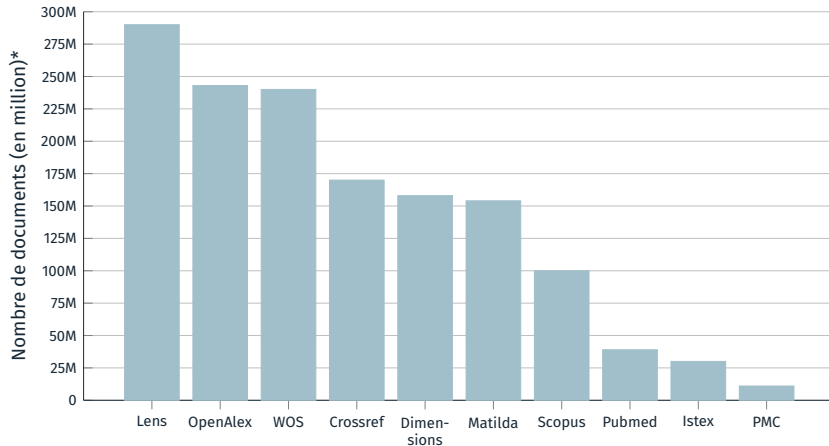


Défis méthodologiques

Istex dans le paysage des réservoirs

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS



Comparaison quantitative des réservoirs (les chiffres peuvent varier en fonction des mises à jour et des méthodologies de comptabilisation)

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Réservoirs	Origine des données	Fournisseurs
Lens	Microsoft Academic, PubMed, Crossref, OpenAlex, Unpaywall, ORCID	Cambia
OpenAlex	Microsoft Academic, Crossref, PubMed, Unpaywall, DOAJ, ORCID, ROR, Internet archive, crawls web, arXiv, Zenodo	OurResearch
Dimensions	Crossref, PubMed, DataCite, arXiv + 130 éditeurs	Digital Science
Matilda	Crossref, PubMed Central, arXiv et RePEc, BASE + Unpaywall, ORCID	I3 & Huma-Num
Crossref	Crossref	Crossref
WOS	Revue sélectionnées	Clarivate
Scopus	Revue sélectionnées	Elsevier
Pubmed	Références sélectionnées (MEDLINE)	National Library of Medicine
PMC	Négociation, OA	National Library of Medicine
Istex	Négociations, Achats, Plan de soutien, GIS CollEX-Persée, OA	Istex

Origine des données (Bouchard, 2024, Gusenbauer, 2022)

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+
Indicateurs biblio.	+	+	+	+	-	-	-

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+
Indicateurs biblio.	+	+	+	+	-	-	-
Indicateurs TDM	-	-	-	-	-	-	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+
Indicateurs biblio.	+	+	+	+	-	-	-
Indicateurs TDM	-	-	-	-	-	-	+
Texte intégral	-	-	-	-	-	+	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+
Indicateurs biblio.	+	+	+	+	-	-	-
Indicateurs TDM	-	-	-	-	-	-	+
Texte intégral	-	-	-	-	-	+	+
Formats TDM standardisés (e.g. TEI)	-	-	-	-	-	-	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

	Lens	OpenAlex	Dimensions	Matilda	Pubmed	PMC	Istex
Filtres	+	+	+	+/-	+	+	+
Tris	+	+	+	+	+	+	+
Enrichissements	+	+	+	+	+	+	++
Recherche assistée	++	+/-	-	-	+	+	++
Formats biblio. (e.g. BIB, MODS)	+	+	+	+	+	+	+
Indicateurs biblio.	+	+	+	+	-	-	-
Indicateurs TDM	-	-	-	-	-	-	+
Texte intégral	-	-	-	-	-	+	+
Formats TDM standardisés (e.g. TEI)	-	-	-	-	-	-	+
Passerelles	-	-	-	-	-	-	+

Propriétés des réservoirs : des usages différents

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources
- ⊕ Des données interoperables, des formats homogénéisés et des données corrigées (moins de prétraitements)

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources
- ⊕ Des données interopérables, des formats homogénéisés et des données corrigées (moins de prétraitements)
- ⊕ Des données enrichies (réocérisation, structuration de texte, métadonnées) : des documents retrouvés et analysés plus facilement

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources
- ⊕ Des données interopérables, des formats homogénéisés et des données corrigées (moins de prétraitements)
- ⊕ Des données enrichies (réocérisation, structuration de texte, métadonnées) : des documents retrouvés et analysés plus facilement
- ⊕ Des millions de textes et de métadonnées téléchargeables facilement

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources
- ⊕ Des données interoperables, des formats homogénéisés et des données corrigées (moins de prétraitements)
- ⊕ Des données enrichies (réocérisation, structuration de texte, métadonnées) : des documents retrouvés et analysés plus facilement
- ⊕ Des millions de textes et de métadonnées téléchargeables facilement
- ⊕ Des connexions vers des outils / plateformes du monde académique

Partie 1 : éléments contextuels | Défis méthodologiques

ISTEX DANS LE PAYSAGE DES RÉSERVOIRS

Istex : infrastructure dédiée à la constitution et à l'analyse de corpus

- ⊕ Des données vérifiées accessibles un seul lieu pour de nombreuses sources
- ⊕ Des données interoperables, des formats homogénéisés et des données corrigées (moins de prétraitements)
- ⊕ Des données enrichies (réocérisation, structuration de texte, métadonnées) : des documents retrouvés et analysés plus facilement
- ⊕ Des millions de textes et de métadonnées téléchargeables facilement
- ⊕ Des connexions vers des outils / plateformes du monde académique
- ⊕ Un cadre juridique sécurisé par une licence appropriée et déjà négociée



Istex



Istex

Qu'est-ce qu'Istex ?

Partie 1 : éléments contextuels | Istex

QU'EST-CE QU'ISTEX?

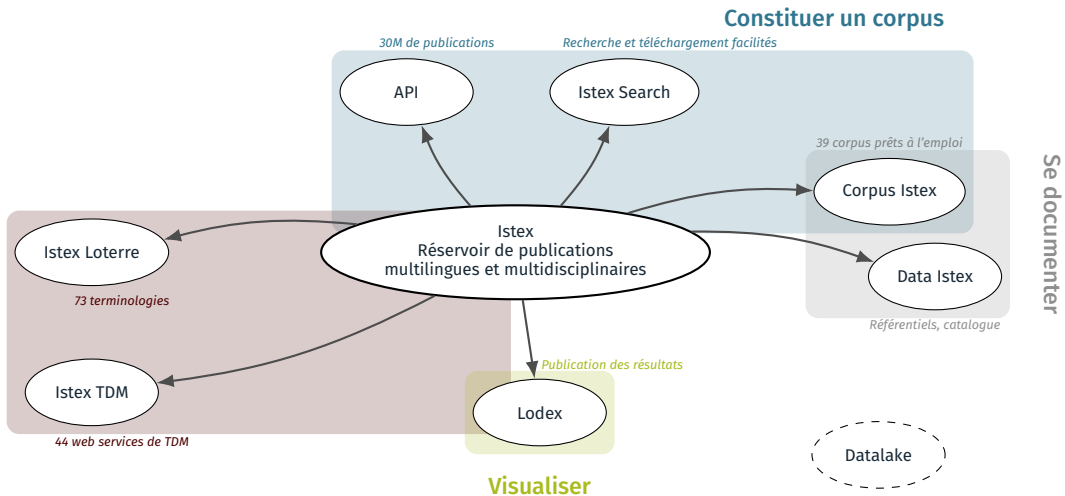
Istex est une infrastructure dédiée à **la constitution et à l'analyse de corpus** (ou fouille de textes, TDM) qui donne accès :

- à des ressources textuelles et lexicales enrichies (publications et terminologies)
- à des outils pour sélectionner et analyser ces ressources

Partie 1 : éléments contextuels | Istex

QU'EST-CE QU'ISTEX?

Annoter, enrichir et analyser





Istex

Pour quels usages ?

Partie 1 : éléments contextuels | Istex

POUR QUELS USAGES?

Une ressource documentaire

- Trouver des articles scientifiques
- Alimenter des bases documentaires, des agrégateurs (Google Scholar, Click & Read, ENT)

Un matériau de recherche

- Constituer des corpus
- Faire de la fouille de textes ou du TDM, du TAL (entraîner des LLM, faire de la RI, etc. e.g. Bénard *et al.*, 2023)

Projet ANR MaTOS



Istex

Chaîne de traitements

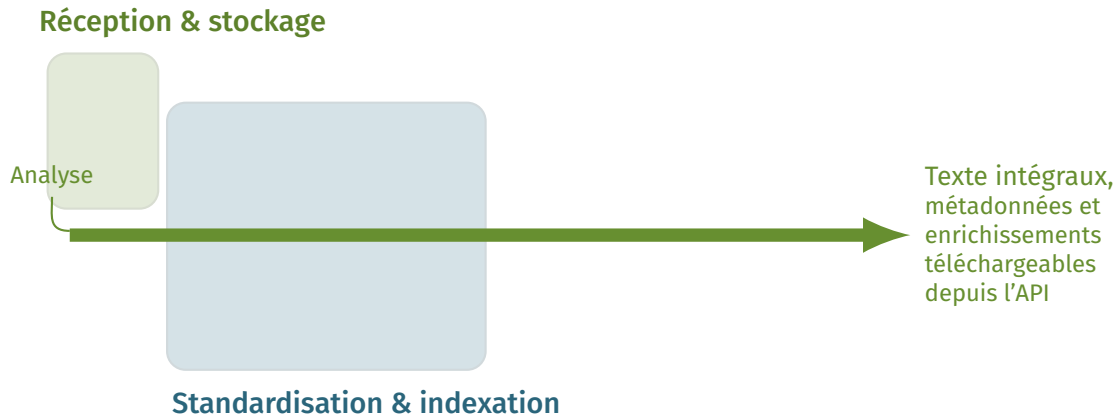
Partie 1 : éléments contextuels | Istex

CHAÎNE DE TRAITEMENTS



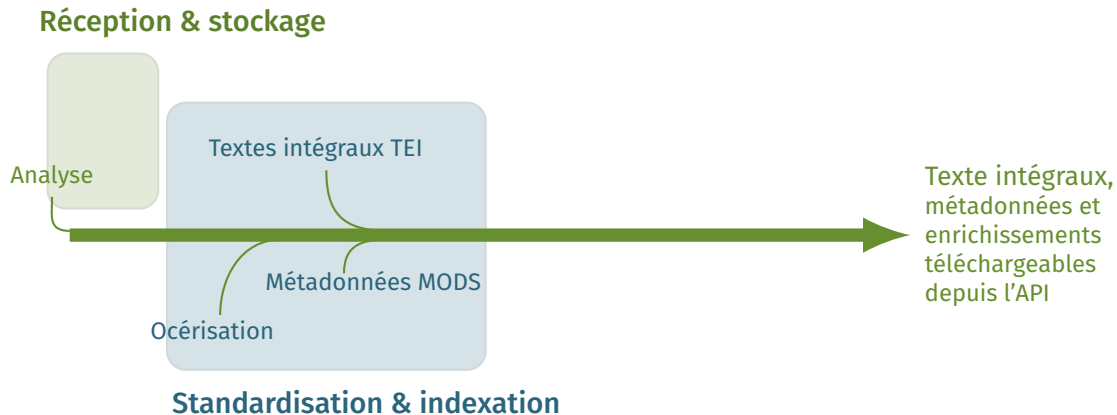
Partie 1 : éléments contextuels | Istex

CHAÎNE DE TRAITEMENTS



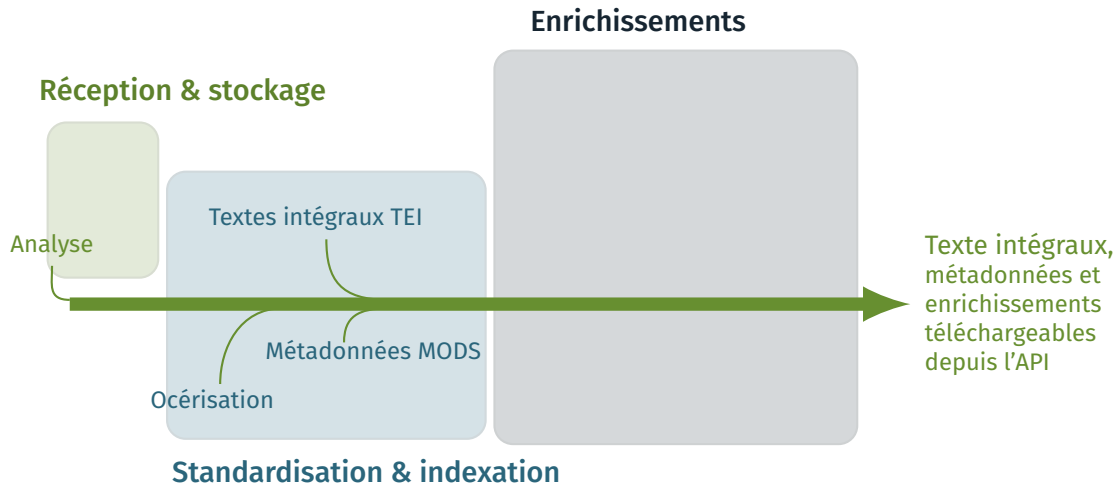
Partie 1 : éléments contextuels | Istex

CHAÎNE DE TRAITEMENTS



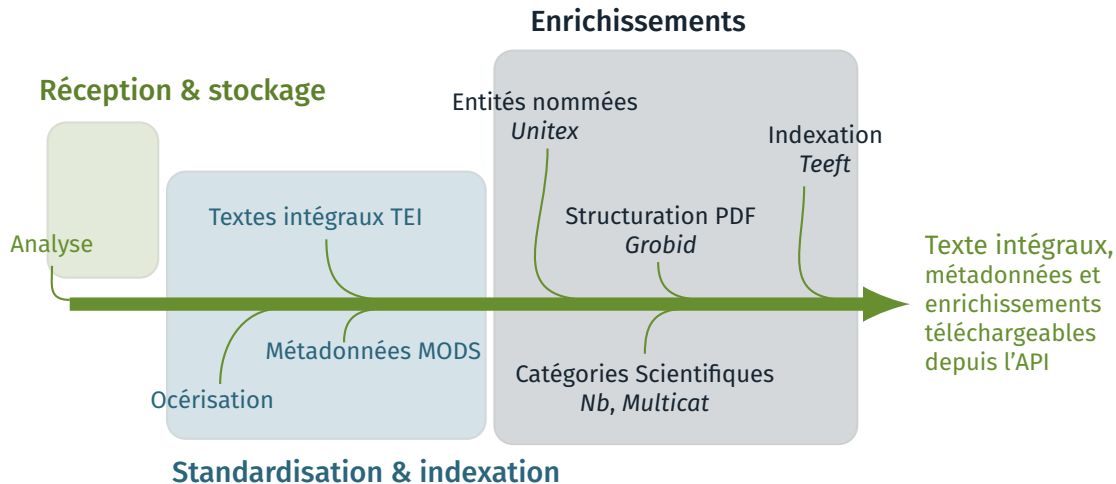
Partie 1 : éléments contextuels | Istex

CHAÎNE DE TRAITEMENTS



Partie 1 : éléments contextuels | Istex

CHAÎNE DE TRAITEMENTS



PARTIE 2 : CONSTITUTION D'UN CORPUS SPÉCIALISÉ



Présentation du cas d'usage

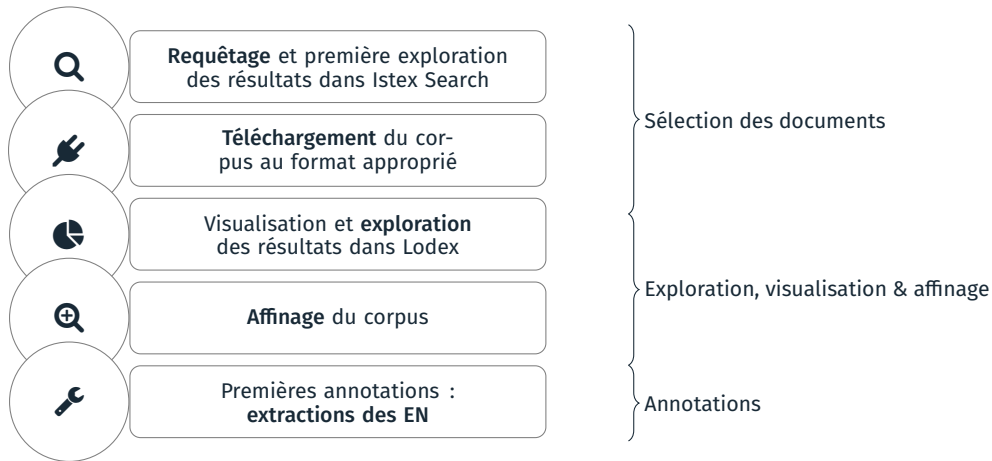
Partie 2 : constitution d'un corpus spécialisé | Présentation du cas d'usage

Objectif Cette recherche en sociologie de l'alimentation vise à comprendre comment les motivations des personnes véganes et végétariennes évoluent au fil du temps (justifications éthiques, sanitaires ou écologiques). L'étude s'intéresse spécifiquement à la manière dont ces transformations sont traitées dans la littérature scientifique internationale.

Moyen Afin d'explorer ces évolutions, un corpus de publications académiques internationales sera constitué et analysé.

Partie 2 : constitution d'un corpus spécialisé | Présentation du cas d'usage

MÉTHODOLOGIE





Sélection des documents



Sélection des documents

Présentation d'Istex Search

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

PRÉSENTATION D'ISTEX SEARCH

Application en ligne qui permet :

- D'interroger l'API grâce à trois modes de requêtage
- De consulter un document dans une notice détaillée
- De naviguer dans les résultats avec différents modes de tri et différents filtres
- De vérifier la compatibilité TDM grâce à des indicateurs techniques
- De sélectionner / exclure des documents pour éliminer simplement le bruit
- De télécharger massivement des documents au format approprié

<https://search.istex.fr>

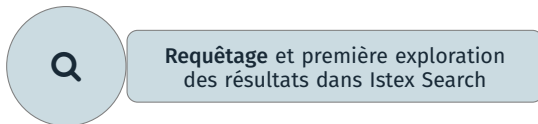


Sélection des documents

Requêtage

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE



Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Requête 1

```
végan vegan véganisme veganism végétarien  
vegetarian végétarisme vegetarianism végétalien  
vegetalian végétalisme vegetalism  
végétarianisme vegetarianism
```

Description Liste de mots-clés anglais et français, coordonnés par l'opérateur booléen OR (= *ou inclusif*), insensibilité à la casse mais sensibilité aux diacritiques (chat = CHAT; éclair ≠ eclair)

Résultats 41 739 documents

Problèmes Il y a du silence; Variantes morphosyntaxiques ou dérivationnelles non recherchées (tokenisation *lacto-végétarien* mais absence de lemmatisation)

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Requête 2

```
végan* vegan* végétar* vegetar* végétalien*  
vegetalian* végétalisme* vegetarianism*
```

Description Recherche des formes morphologiquement apparentées grâce à la troncature symbolisée par *

Résultats 53 594 documents

Problèmes Il y a du bruit (documents récents)

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTE

Original Article

Effect of slope on the forest structure of the Atlantic Forest domain in southern Brazil

Efeito da inclinação na estrutura de uma floresta do domínio Mata Atlântica no sul do Brasil

D. C. Souza^{a*} , L. R. Souza^b , E. V. Couto^a , M. G. Caxambú^a  and A. P. Peron^{a,c} 

^aUniversidade Tecnológica Federal do Paraná – UTFPR, Departamento de Biodiversidade e Conservação da Natureza – DABIC, 81.531-980, Curitiba, PR, Brasil

^bUniversidade Tecnológica Federal do Paraná – UTFPR, Curso de Engenharia Ambiental, Campo Mourão, PR, Brasil

^cUniversidade Tecnológica Federal do Paraná – UTFPR, Programa de Pós-graduação em Inovações Tecnológicas – PPGIT, Campo Mourão, PR, Brasil

Effet de la pente sur la structure d'une forêt brésilienne

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTE

Original Article

Effect of slope on the forest structure of the Atlantic Forest domain in southern Brazil

Efeito da inclinação na estrutura de uma floresta do domínio Mata Atlântica no sul do Brasil

D. C. Souza^{a*}, L. R. Souza^b, E. V. Couto^a, M. G. Caxambú^a and A. P. Peron^{a,c}

^aUniversidade Tecnológica Federal do Paraná – UTFPR, Departamento de Biodiversidade e Conservação da Natureza – DABIC, Curitiba, PR, Brasil

^bUniversidade Tecnológica Federal do Paraná – UTFPR, Curso de Engenharia Ambiental, Campo Mourão, PR, Brasil

^cUniversidade Tecnológica Federal do Paraná – UTFPR, Programa de Pós-graduação em Inovações Tecnológicas – PPGIT, Campo Mourão, PR, Brasil

Effet de la pente sur la structure d'une forêt brésilienne

the temporary crop begins, being these collections every 20 meters along the slope. In these samples, macro- and micro-nutrients were analyzed.

The physical analysis of the remaining soil was carried out through three trenches called T1, T2 and T3; T1 located at high slope, and T2 and T3 at medium slope. Soil particle size analyses were made for each trench, for soil classification according to the Brazilian Soil Classification System (Embrapa, 2006). In the trenches, color and horizon were determined in situ. As for the color, the Munsell Color Chart was used as a reference (USDA, 1988). The slope of the area was determined along a transect that covered all the plots, using GNSS spectra promark 220, accurate to 0.06m. The same equipment was used to fix the four vertices of the plots, so in future studies the same plots can be used.

2.3. Phytosociology

The forest structure was sampled in 12 continuous rectangular plots, with 30 X 20 m, totaling 0.72 ha.

2.4. Data analysis

Species were analyzed for clusters according to their distribution in the area by means of a Detrended Correspondence Analysis (DCA). These distribution data were used together with soil chemical properties to set some pattern between the plots. These were analyzed by Canonical Correspondence Analysis (CCA), which allows associating species distribution with abiotic variables. Pedological variables with the greatest pressure on phytosociological characteristics were tested for patterns in the environmental gradient by Regression Analysis to identify the influence of the variable and creating prediction models. All analyses were run in the R program. To test normality and perform regression analysis, the Shapiro test function and the lm function of the R base were used (R Core Team, 2020). For multivariate analysis, the decorana and cca functions of the **vegan** package were used (Oksanen et al., 2020).

3. Results and Discussion

Vegan est le nom d'un package R cité dans le corps du texte

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Research paper

Preventing intense false positive and negative reactions attributed to the principle of ELISA to re-investigate antibody studies in autoimmune diseases

Kuniaki Terato^{a,*}, Christopher T. Do^a, Dawn Cutler^a, Takaki Waritani^a, Hiroshi

^a Chondrex Inc., 2607 151st Place NE, Redmond, WA 98052, United States

Publication sur les maladies auto-immunes

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Research paper

Preventing intense false positive and negative results attributed to the principle of ELISA to re-investigate studies in autoimmune diseases

Kuniaki Terato^{a,*}, Christopher T. Do^a, Dawn Cutler^a, Takaki W

^a Chondrex Inc., 2607 151st Place NE, Redmond, WA 98052, United States

Publication sur les maladies auto-immunes

- Kijima, Y., Iwatsuki, S., Akamatsu, H., Terato, K., Kuwabara, Y., Ueda, S., Shionoya, H., 2009. Natural antibodies to pathogenic bacteria and their toxins in whey protein concentrate. *Food Sci. Technol. Res.* 56, 475.
- Kjeldsen-Kragh, J., 1999. Rheumatoid arthritis treated with vegetarian diets. *Am. J. Clin. Nutr.* 70, 594S.
- Komiya, I., Arimura, Y., Nakabayashi, K., Yamada, A., Osaki, T., Yamaguchi, H., Kamiya, S., 2011. Increased concentrations of antibody against heat shock protein in patients with myeloperoxidase anti-neutrophil cytoplasmic autoantibody positive microscopic polyangiitis. *Microbiol. Immunol.* 55, 531.
- Ligier, S., Fortin, P., Newkirk, M., 1998. A new antibody in rheumatoid arthritis targeting glycated IgG: IgM anti-IgG-AGE. *Br. J. Rheumatol.* 37, 1307.
- Lucke, K., Miehke, S., Jacobs, E., Schuppler, M., 2006. Prevalence of *Bacteroides* and *Prevotella* spp. in ulcerative colitis. *J. Med. Microbiol.* 55, 617.
- Maes, M., Twisk, F., Kubera, M., Ringel, K., Leunis, J., Geffard, M., 2011. Increased IgA responses to the LPS of commensal bacteria is associated with inflammation and activation of cell-mediated immunity in chronic fatigue syndrome. *J. Affect. Disord.* 136, 909.

Terme présent dans les références bibliographiques

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Requête 3

```
title:(végan* vegan* végétar* vegetar*  
végétalien* vegetarian* végétalisme*  
vegetalism*) abstract:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarian*)  
subject.value:(végan* vegan* végétar*  
vegetar* végétalien* vegetarian*  
végétalisme* vegetarian*)  
keywords.teeft:(végan* vegan*  
végétar* vegetar* végétalien*  
vegetarian* végétalisme* vegetarian*)
```

Description Cibler les champs les plus à même de renvoyer des documents pertinents : `title`, `abstract`, `subject.value`, `keywords.teeft`

Résultats 8 143 documents

Problèmes Il y a du bruit (documents anciens)

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatynge of flesshe, and lyke constitutyons of men. &c

An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatynge of flesshe, and lyke constitutyons of men. &c

Erasmus, Desiderius d. 1536

Pas de résumé

✕ REVENIR AUX RÉSULTATS

➦ PARTAGER LE DOCUMENT

Informations sur ce document

database other eebo 1534

ark:/67375/6ZK-FG886LQR-Q

Consulter ce document

Texte intégral

tiff pdf zip

Métadonnées

xml mods json

☐ SÉLECTIONNER

☒ EXCLURE

Publication de 1534

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTE

An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatyng of flesshe, and lyke constitutyons of men. &c

An epystell of ye famous doctor Erasmus of Roterdam : vnto the reuerende father & excellent prince, Christofer bysshop of Basyle, co[n]cernyng the forbedyng of eatyng of flesshe, and lyke constitutyons of men. &c

Erasmus, Desiderius d. 1536

Pas de résumé

✕ REVENIR AUX RÉSULTATS

➤ PARTAGER LE DOCUMENT

Informations sur le document

database | oti

ark:/67375/62K

Consulter ce document

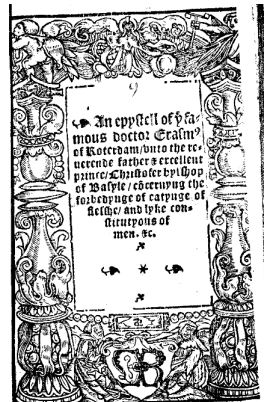
Texte intégral

tiff pdf zip

Métadonnées

xml mods json

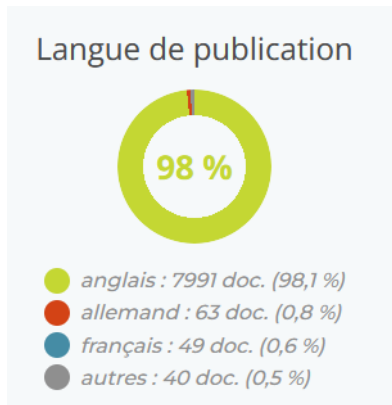
Publication de 1534



Document sans résumé, non océrisé

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

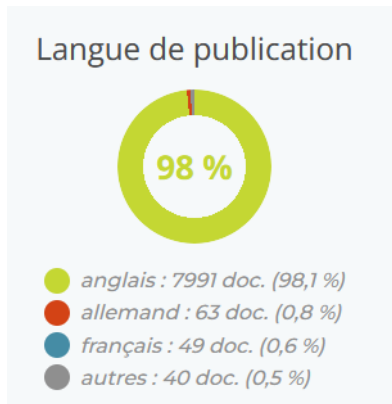
REQUÊTAGE



Indicateur des langues de publication

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE



Indicateur des langues de publication

404

Жива Бенчик

В анализе короткого рассказа Евгения Попова 'Как съели петуха' мы могли бы прибегнуть к интересному положению Клода Леви-Стросса, согласно которому кухня любого общества представляет собой особый язык, на котором оно бессознательно раскрывает свою структуру (Lévi-Strauss 1983: 373). Дело в том, что, по утверждению французского ученого, кулинарные приемы, как и язык, можно анализировать в конститутивных элементах, так называемых густемах, организованных по определенным структурам оппозиции и корреляции. Так, например, английскую кухню можно отличать от французской посредством трех оппозиций:

эндогенное/экзогенное (то есть национальное или заграничное сырье); центральное/периферийное (главное блюдо и сопутствующие блюда); выраженное/невыраженное (то есть вкусное или безвкусное) [...]. Другими словами, в английской кухне на основные, безвкусно приготовленные блюда, идут отечественные продукты; а на сопутствующие – импортная продукция, в которой все дифференциальные качества ярко выражены (чай, фруктовый пирог, апельсиновый мармелад, портвейн). И наоборот, во французской кухне оппозиция эндогенное/экзогенное значительно ослаблена или отсутствует, а одинаково выраженные густемы сочетаются друг с другом, как в центральной, так и в периферийной кухне (Lévi-Strauss 1983: 373).

Document russe

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

REQUÊTAGE

Requête 4

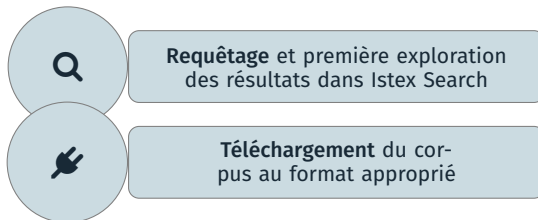
```
(title:(végan* vegan* végétar* vegetarian*  
végétalien* vegetarian* végétalisme*  
vegetalism*) abstract:(végan* vegan*  
végétar* vegetarian* végétalien*  
vegetarian* végétalisme* vegetarian*)  
subject.value:(végan* vegan* végétar*  
vegetar* végétalien* vegetarian*  
végétalisme* vegetarian*))  
keywords.teeft:(végan* vegan*  
végétar* vegetarian* végétalien*  
vegetarian* végétalisme* vegetarian*))  
AND publicationDate:[1950 TO 2025]  
AND abstract:* AND language:(eng fre)
```

Description Contraindre la présence d'un résumé, contraindre les langues, cibler les dates de publications grâce à l'intervalle [1950 TO *] le tout grâce au parenthésage et à l'opérateur AND

Résultats 4 018 documents

Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

TÉLÉCHARGEMENT



Partie 2 : constitution d'un corpus spécialisé | Sélection des documents

TÉLÉCHARGEMENT

Configurez votre téléchargement

USAGE PERSONNALISÉLODEXCORTEXTGARGANTEXTNOOJ

☐ Texte intégral
☐ PDF
☐ TEI
☐ TXT
☐ CLEANED
☐ ZIP
☐ TIFF

☒ Métadonnées
☒ JSON
☐ XML
☐ MOOS
☐ Annexes
☐ Couvertures

☐ Enrichissements
☐ multicat
☐ nb
☐ grobidFulltext
☐ refBibs
☐ teelt
☐ unitex

Trier par : pertinence & qualité ▾

Télécharger 4 018 / 4 018 TOUT

Format de l'archive : ZIP ▾ Compression : moyenne ▾

TÉLÉCHARGER

Lodex

Application web open-source dédiée aux données structurées qui permet de visualiser et d'enrichir ses données puis de les transformer en site web.

[EN SAVOIR PLUS](#)

Requête

```
(title:(végan* vegan* végétar* végétar* végétalien* végétalien* végétalisme* végétalisme*) abstract:(végan* vegan* végétar* végétar* végétalien* végétalien* végétalisme* végétalisme*) subject:vegetar* value:(végan* vegan* végétar* végétar* végétalien* végétalien* végétalisme* végétalisme*)
```

Requête brute complète

```
https://api.istex.fr/document?q=(title:(végan* vegan* végétar* végétar* végétalien* végétalien* végétalisme* végétalisme*) abstract:(végan* vegan* végétar* végétar* végétalien* végétalien* végétalisme* végétalisme*)
```

Modale de téléchargement des données dans Istex Search



Exploration, visualisation & affinage

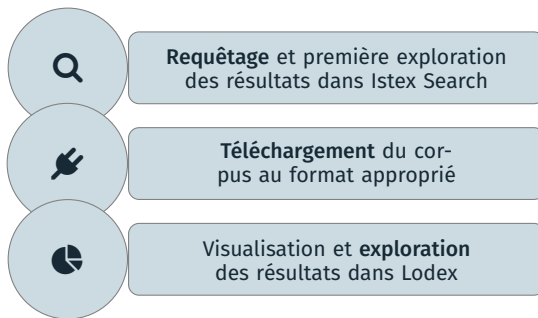


Exploration, visualisation & affinage

Présentation de Lodex

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEx

« *Linked Open Data EXperiment* »

Application web open-source dédiée à la **visualisation** de données structurées
(Gregorio *et al.*, 2019)

<https://www.lodex.fr>

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web À partir de différents formats (CSV, JSON, TXT,...)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web

À partir de différents formats (CSV, JSON, TXT,...)



Explorer ses données

À l'aide de différents graphiques, filtres et au travers de données complémentaires

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web

À partir de différents formats (CSV, JSON, TXT,...)



Explorer ses données

À l'aide de différents graphiques, filtres et au travers de données complémentaires



Enrichir ses données

Grâce à un catalogue de web services

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web

À partir de différents formats (CSV, JSON, TXT,...)



Explorer ses données

À l'aide de différents graphiques, filtres et au travers de données complémentaires



Enrichir ses données

Grâce à un catalogue de web services



Aligner ses données

Avec des données similaires ou connexes (requête SPARQL)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web

À partir de différents formats (CSV, JSON, TXT,...)



Explorer ses données

À l'aide de différents graphiques, filtres et au travers de données complémentaires



Enrichir ses données

Grâce à un catalogue de web services



Aligner ses données

Avec des données similaires ou connexes (requête SPARQL)



Exporter ses données

Totalement ou partiellement (JSONL, CSV...)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Transformer ses données en site web

À partir de différents formats (CSV, JSON, TXT,...)



Explorer ses données

À l'aide de différents graphiques, filtres et au travers de données complémentaires



Enrichir ses données

Grâce à un catalogue de web services



Aligner ses données

Avec des données similaires ou connexes (requête SPARQL)



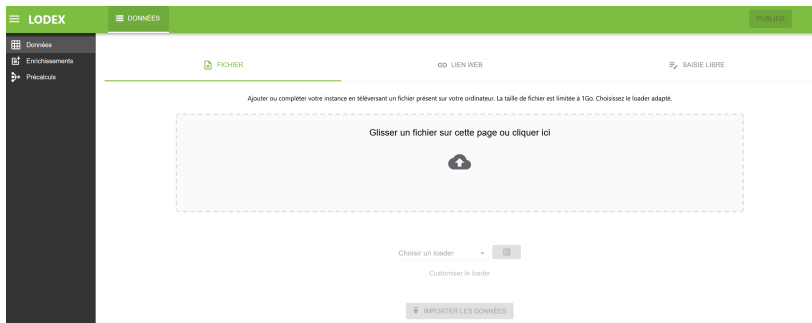
Exporter ses données

Totalement ou partiellement (JSONL, CSV...)

Exemple

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX



Interface d'import des données dans Lodox

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX

LODEX									
DONNÉES AFFICHAGE ANNOTATIONS									
COLONNES FILTRES DENSITÉ AJOUTER									
☐	uri	Titre	Auteur(s)	Affiliation(s)	Revue ou monogra...	Auteur(s) monogra...	ISSN	e-ISSN	Volume
☐	"ark:/67375/6H6-4D6"	"Végétarisme, végétal	["Sasha Mathieu","Gér	["Université Sorbonne	"La Presse Médicale"		"0755-4982"	undefined	"45"
☐	"ark:/67375/6S2-BC7"	"Conocimientos alime	["Jerusa Brignardello		"Revista chilena de nu		undefined	"0717-7518"	"40"
☐	"ark:/67375/6S2-JRZ"	"Relatos de vida que	["Carolina Díaz","Leny		"Revista chilena de nu		undefined	"0717-7518"	"49"
☐	"ark:/67375/VQC-612"	"Do vegetarians have	["Susan A. New"]	["Centre for Nutrition	"Osteoporosis Internat		"0937-941X"	"1433-2965"	"15"
☐	"ark:/67375/6H6-2M0"	"Exposure to endocrin	["Karen Tordjman","La	["Institute of Endocrin	"Environment Internati		"0160-4120"	undefined	"97"
☐	"ark:/67375/6H6-CHF"	"High compliance with	["Jakub G. Sobiecki","	["Cancer Epidemiolog	"Nutrition Research"		"0271-5317"	undefined	"36"
☐	"ark:/67375/2M2-PKW"	"Erythromycin-resista	["Vesna Milanović","Ar	["Dipartimento di Sci	"PLOS ONE"		undefined	"1932-6203"	undefir

Transformation des données par le loader

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

PRÉSENTATION DE LODEX

The screenshot displays the LODEX web application interface. At the top, a green navigation bar contains the 'LODEX' logo, a menu icon, and tabs for 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS'. A 'DÉPUBLIER' button is located on the right. A dark 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 'DONNÉES PUBLIÉES'. It features two input fields: 'DATASET - Titre' with the value '[qMKk]' and 'DATASET - Description' with the value 'Description du corpus'. A '+ NOUVEAU CHAMP' button is on the right. Below these fields are two rows of data, each with a checkbox, a label, and a home icon. The first row has a checkbox, the label '[qMKk]', and the title 'Titre'. The second row has a checkbox, the label '[CYeG] Description du corpus', and a home icon.

Import du modèle



Exploration, visualisation & affinage

Exploration

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

EXPLORATION



Corpus Véganisme et Végétarisme

Description du corpus

Le présent travail s'appuie sur le **corpus « Véganisme et végétarisme »**, conçu comme ressource documentaire destinée à l'analyse des transformations sociales et culturelles liées aux régimes alimentaires alternatifs. Ce corpus s'inscrit dans une perspective de **sociologie de l'alimentation**, en mettant l'accent sur l'évolution des représentations, des pratiques et des aspirations associées au végétarisme et au végétarisme.

Constitué à partir des publications indexées dans la base **Istex Search**, il bénéficie d'un enrichissement automatisé grâce à plusieurs services de **text and data mining (TDM)**, notamment :

- **Teefit**, qui assure l'extraction de mots-clés et facilite l'identification des thématiques dominantes ;
- **entityTag**, qui procède à la reconnaissance des entités nommées (individus, lieux, organismes.).

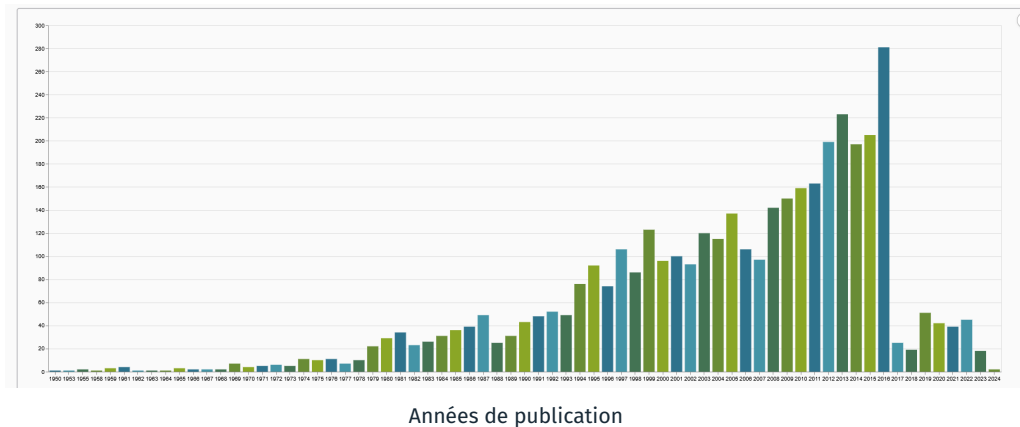
Ce dispositif permet non seulement de cartographier la production scientifique consacrée à ces pratiques alimentaires, mais également d'envisager une analyse diachronique des discours académiques et de leurs évolutions. L'objectif est ainsi de fournir un matériau robuste et interdisciplinaire, apte à nourrir une réflexion sur les dynamiques sociales, culturelles et politiques qui traversent les débats contemporains autour du végétarisme et du végétarisme.

[CONSULTER LA REQUÊTE INITIALE](#)

Site web du corpus créé avec Lodex

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

EXPLORATION

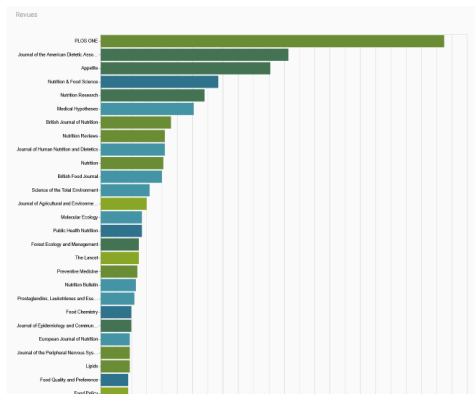


EXPLORATION



Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

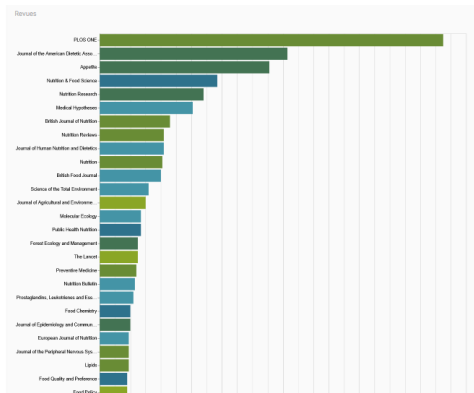
EXPLORATION



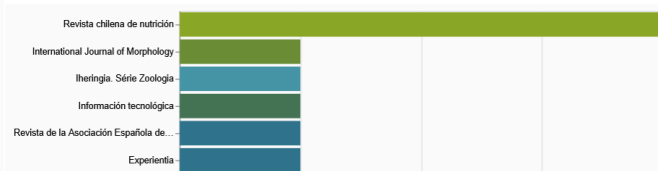
Revues du corpus

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

EXPLORATION



Revue du corpus



Revue hispanophones (gestion du multilinguisme)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

EXPLORATION

Rev Chil Nutr 2020; 47(5): 782-791.

<http://dx.doi.org/10.4067/S0717-75182020000500782>

Artículo Original / Original Article

Estado nutricional, hábitos de alimentación y de estilo de vida en vegetarianos de Asunción y Gran Asunción, Paraguay

Nutritional status, eating habits and lifestyle factors among vegetarians from Asunción and Great Asunción, Paraguay

RESUMEN

Una dieta vegetariana se define por la exclusión total o parcial de alimentos de origen animal. Cuando es planificada adecuadamente puede prevenir y tratar enfermedades. Sin embargo, una dieta restrictiva en ciertos alimentos debe ser evaluada. El propósito de esta investigación fue describir parámetros antropométricos, hábitos de alimentación y de estilo de vida en ovo-lácteo-vegetarianos y veganos y explorar diferencias en la ingesta de nutrientes críticos entre ambos grupos. Estudio descriptivo de corte transversal, realizado en ovo-lácteo-vegetarianos y veganos de Asunción, Paraguay.

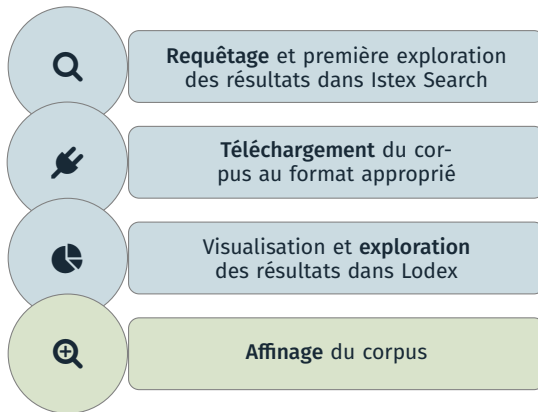
Meliessa Penner Teichgräb¹, Natalia Elizabeth González Cañete^{1*}.

¹. Carrera de Nutrición. Facultad de Ciencias Médicas.
Universidad del Pacífico. Asunción, Paraguay.

Exemple d'article issu de la revue *Revista chilena de nutrición*

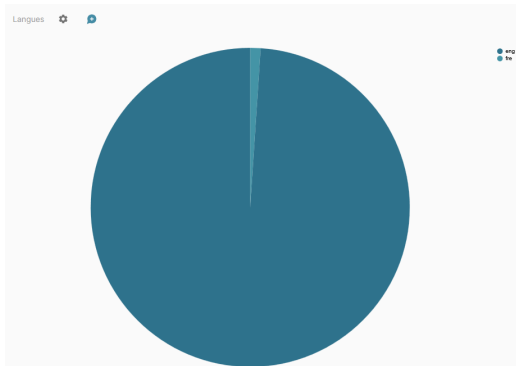
Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

AFFINAGE



Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

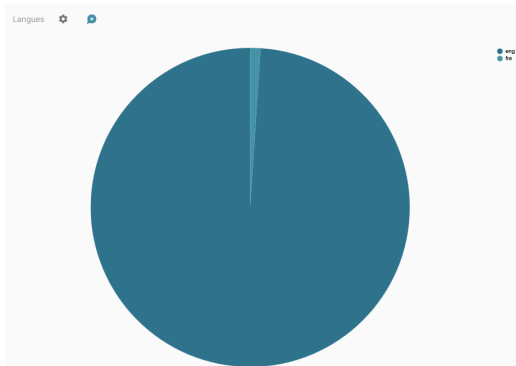
AFFINAGE



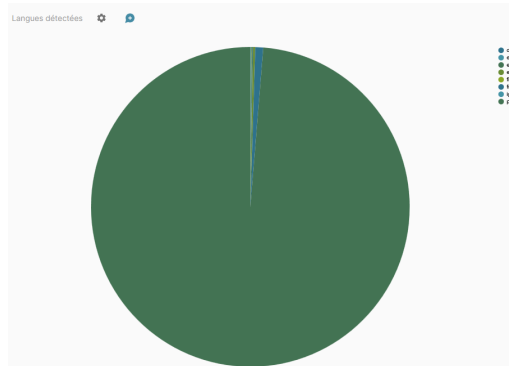
Langues : français (42), anglais (3 976)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

AFFINAGE



Langues : français (42), anglais (3 976)



Langues issues du WS (8 langues; 3 955 en/4 018)

Partie 2 : constitution d'un corpus spécialisé | Exploration, visualisation & affinage

AFFINAGE

The screenshot shows the LODEX interface with a green header bar containing the menu icon, 'LODEX', and tabs for 'DONNÉES', 'AFFICHAGE', and 'ANNOTATIONS'. A 'DÉPUBLIER' button is on the right. A left sidebar lists 'Données', 'Enrichissements', 'Précalculs', and 'Ressources cachées'. The main area has tabs for 'COLONNES', 'FILTRES', 'DENSITÉ', and 'AJOUTER'. A table of documents is displayed with columns: uri, Titre, Auteur(s), Affiliation(s), Revue ou monogra..., Auteur(s) monogra..., ISSN, e-ISSN, and Volume. A filter dropdown menu is open over the first row, showing 'Colonnes', 'Opérateur', and 'Valeur'. The filter is set to 'Détection de la la' with the operator 'contient' and the value 'es'. A button 'SUPPRIMER LE(S) LIGNE(S) FILTRÉE(S)' is at the top right of the table.

uri	Titre	Auteur(s)	Affiliation(s)	Revue ou monogra...	Auteur(s) monogra...	ISSN	e-ISSN	Volume
				"Revista chilena de nu. []		undefined	"0717-7518"	"40"
	"ark:/67375/6S2-JRZx" "Relatos de vida que e	["Carolina Díaz", "Lery"]		"Revista chilena de nu. []		undefined	"0717-7518"	"49"
	"ark:/67375/6S2-HNLj" "La Transición a la Ag	["Natalia Morales", "Vh"]		"International Journal []		undefined	"0717-9502"	"36"
	"ark:/67375/6S2-P2Cf" "Estado nutricional, hñ	["Melissa Penner Tei"]		"Revista chilena de nu. []		undefined	"0717-7518"	"47"
	"ark:/67375/6S2-10L3" "Patrón de dieta pro-v	["Miguel Ángel Martín"]		"Revista chilena de nu. []		undefined	"0717-7518"	"41"

Suppression des documents dans le dataset



Premières annotations : extractions d'EN

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

PRÉSENTATION D'ISTEX TDM





Premières annotations : extractions d'EN

Présentation d'Istex TDM

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

Les services de fouille de textes

- Istex met a disposition 44 web services dans Istex TDM (Cuxac, 2024)
 - Prétraitement (*textExtract*)
 - Classification (*aiAbstractCheck*)
 - Validation (*bibCheck*)
 - Extraction (*diseaseTag*)
 - Alignement (*loterre-resolvers*)
- Un web service est un **programme mono-tâche, frugal avec un paramétrage minimal**
- Les web services peuvent être utilisés dans Lodex, via TDM Factory, ou en ligne de commande



Premières annotations : extractions d'EN

Extraction d'information

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION

LODEX									
DONNÉES AFFICHAGE ANNOTATIONS									
DÉPUBLIER									
COLONNES FILTRES DENSITÉ AJOUTER									
ence...	Catégories INIST	Mots-clés (teeff)	Entités nommées (...)	Types d'enrichisse...	Score qualité	Version PDF	Nb mots PDF	Nb mots par page ...	Nb caractères PDF
heral &	[["Nom": "3 - sciences	undefined	undefined	[["multicat", "nb"]	9.1	1.7	6181	773	36376
doctrinc	[["Nom": "3 - sciences	[["nutr", "bone health", \	[["date": ["1960s"], "org	[["grobidFulltext", "multi	8.667	1.4	8176	818	43570
vironm	[["Nom": "4 - toxicologi	[["aminim", "lbrms", "ph	undefined	[["grobidFulltext", "multi	10	1.7	8185	1023	45294
trillon &	[["Nom": "3 - sciences	[["vegan", "eater", "meal	undefined	[["multicat", "nb", "teeff"]	10	1.7	11377	813	58802
	[["Nom": "4 - microbiolo	[["omnivore", "vegan", "s	undefined	[["nb", "teeff"]	9.964	1.6	8087	539	45642
trillon &	[["Nom": "1 - sciences	[["vegetarian", "vegan",	undefined	[["grobidFulltext", "multi	9.64	1.7	8833	981	47920
	[["Nom": "1 - sciences	undefined	undefined	[["nb"]	7.461	1.5	8917	495	48408
	[["Nom": "3 - sciences	[["vitamina", "nutr", "diet	undefined	[["nb", "teeff"]	8.121	1.5	6583	823	35651
trillon &	[["Nom": "2 - philosoph	[["vegan", "masculinity",	undefined	[["grobidFulltext", "multi	9.568	1.7	8983	1123	49362

Istex contient des annotations mais elles sont parfois incomplètes (Unitex)

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION

The screenshot shows the LODEX web interface. The top navigation bar is green with the LODEX logo and three tabs: DONNÉES (selected), AFFICHAGE, and ANNOTATIONS. A left sidebar contains four menu items: Données, Enrichissements, Précalculs, and Ressources cachées. The main content area has a form for entity extraction. It includes a 'Nom' field with the value 'EN extraction' and a 'LANCER' button. Below this, it shows '0 erreurs' and a 'RÉESSAYER' button. The 'Statut' is 'Non démarré' with a 'VOIR LES LOGS' link. There is a 'Mode avancé' toggle switch. The 'URL du web service' field contains 'https://person-ner.services.istex.fr/v1/tagger' with a small green icon button. Below the URL are two dropdown menus: 'Colonne de la source' (set to 'Résumé') and 'Sous-chemin'. At the bottom are three buttons: 'SUPPRIMER' (orange), 'ANNULER', and 'SAUVEGARDER' (green).

LODEX DONNÉES AFFICHAGE ANNOTATIONS

Données Enrichissements Précalculs Ressources cachées

Nom **▶ LANCER**

0 erreurs **RÉESSAYER**

Statut : Non démarré [VOIR LES LOGS](#)

☐ Mode avancé

URL du web service 

Colonne de la source

SUPPRIMER **ANNULER** **SAUVEGARDER**

Extraction des entités nommées de lieux, organismes et personnes grâce au web service multilingue *entitytag*

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION

Colonne **EN extraction** pour la ligne **ark:/67375/G14-FS05MDP9-V**

```
"root" : { 3 items
  ▶ "PER" : [] 0 items
  ▼ "LOC" : [ 2 items
    0 : string "South Africa"
    1 : string "South Africa"
  ]
  ▶ "ORG" : [] 0 items
}
```

ANNULER



SAUVEGARDER



Résultats de la détection des entités nommées

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION

Grand-Duché de Luxembourg



id: grandduchedeluxembourg

cartographyCode: LUX

about: <http://data.loterre.fr/ark:/67375/9SD-HXXBRCFQ-F>

prefLabel@fr: Luxembourg

prefLabel@en: Luxembourg

wikidataURI : <https://www.wikidata.org/wiki/Q1842>

geonameURI: <https://www.geonames.org/2960313>

countryCode: LU

latitude: 49.765705074151

longitude: 5.965223432344

localization@en: Western Europe

localization@fr: Europe occidentale

Traitement des entités nommées de lieux : utilisation du web service *Pays et subdivision*

Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION

Colonne **Loterre** pour la ligne **ark:/67375/G14-FS05MDP9**

V

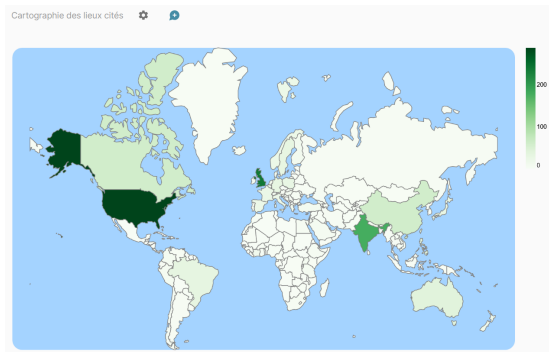
```

"root" : [ 2 items
  0 : { 11 items
    "cartographyCode" : string "ZAF"
    "about" :
      string "http://data.loterre.fr/ark:/67375/9SD-W45W8BVX-4"
    "prefLabel@fr" : string "Afrique du Sud"
    "prefLabel@en" : string "South Africa"
    "wikidataURI" : string "https://www.wikidata.org/wiki/Q3926"
    "geonameURI" : string "https://www.geonames.org/953987"
    "countryCode" : string "ZA"
    "latitude" : string "-28.9621061911244"
    "longitude" : string "25.1173962123943"
    "localization@en" : [] 0 items
    "localization@fr" : [] 0 items
  }
]
```

Résultats de l'alignement avec Loterre

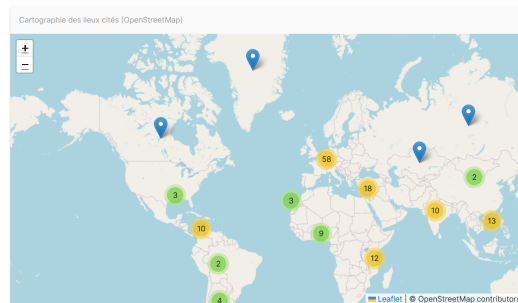
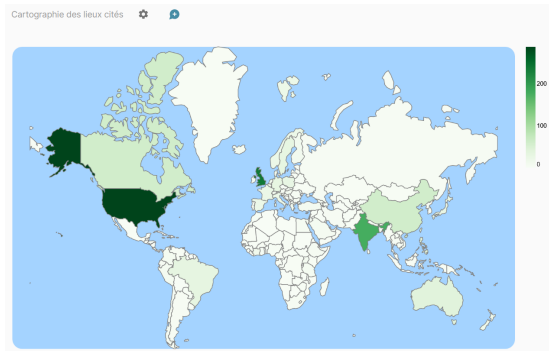
Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION



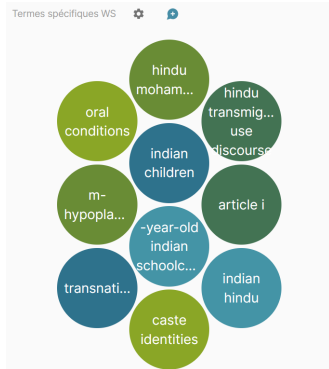
Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION



Partie 2 : constitution d'un corpus spécialisé | Premières annotations : extractions d'entités nommées

EXTRACTION D'INFORMATION



Rapport entre les mots-clés et pays cités

'I don't eat meat': Discourse on food among transnational Hindus

Jennifer B. Saunders

Hindu transmigrants use discourse on diet as a way to maintain connections with India, as well as to construct Indian, Hindu and caste identities. In this article, I argue that such discourse on food is a meta-discourse that reframes the symbolic meaning of food in the transnational context. This article examines a transnational Hindu community's discourse on food, and pairs R.S. Khare's arguments about the communicative function of food in a South Asian context with transnational and performance theories, as well as with Arjun Appadurai's argument about the significance of imagination in creating lived realities. Through their narratives involving food, this community is actively engaged in shifting the meanings of what it eats to emphasise their connections with each other, and with India. Thus, a vegetarian diet and the use of 'authentic' Indian ingredients become the symbols of Indian identity through discourse, which is then solidified through the acts of cooking and eating. This article is based on fieldwork conducted with an extended transnational Hindu

Exemple d'article en sociologie



Conclusion

Conclusion

Une méthodologie itérative : examen des données ↔ révision de la requête

- La pertinence scientifique : circonscrire au maximum le sujet quitte à créer du silence?

Conclusion

Une méthodologie itérative : examen des données ↔ révision de la requête

- La pertinence scientifique : circonscrire au maximum le sujet quitte à créer du silence ?
- La pertinence technique : se limiter aux documents possédant un résumé, un PDF texte, un texte nettoyé, un nombre de mots / pages raisonnable ?

Conclusion

Une méthodologie itérative : examen des données ↔ révision de la requête

- La pertinence scientifique : circonscrire au maximum le sujet quitte à créer du silence ?
- La pertinence technique : se limiter aux documents possédant un résumé, un PDF texte, un texte nettoyé, un nombre de mots / pages raisonnable ?

Des aides précieuses développées dans Istex

- L'interface Istex Search

Conclusion

Une méthodologie itérative : examen des données ↔ révision de la requête

- La pertinence scientifique : circonscrire au maximum le sujet quitte à créer du silence ?
- La pertinence technique : se limiter aux documents possédant un résumé, un PDF texte, un texte nettoyé, un nombre de mots / pages raisonnable ?

Des aides précieuses développées dans Istex

- L'interface Istex Search
- L'appui de Lodex et des web services



Annexes et références

Les outils Istex



Istex

Site général, vitrine du possible

www.istex.fr



Data

Informations sur les données Istex (collections, enrichissements) & accès aux corpus (Cuxac *et al.*, 2019)

data.istex.fr



Search

Application pour interroger l'API et télécharger des corpus

search.istex.fr



Lodex

Outil de création de sites web (visualisation / enrichissement de données) (Gregorio *et al.*, 2019)

www.lodex.fr



TDM

Bibliothèque de web services (Cuxac, 2024)

services.istex.fr



**Lo-
terre**

Bibliothèque de terminologies multilingues (Khayari *et al.*, 2021)

www.loterre.fr

Documentation & tutoriels

Se documenter

- Documentation Istex : `https://doc.istex.fr/tdm`
- Documentation API Istex : `https://doc.istex.fr/api`
- Documentation Lodex : `https://www.lodex.fr/docs/documentation/`

S'authentifier

- Vérifier ses droits d'accès : `https://api.istex.fr/auth`

Informations & contacts

Se tenir informé

- Article d'actualité : <https://www.istex.fr/category/actualites/>

Chercher de l'aide / Contribuer à l'amélioration

- Via le formulaire : <https://www.istex.fr/contact/>
- Via les listes : contact@listes.istex.fr,
https://groupes.renater.fr/sympa/subscribe/lodex?previous_action=info

Références

- BENZITOUN, C. et CAPPEAU, P. (2025). Les corpus et leur exploitation. *Encyclopédie Grammaticale du Français*, En ligne. Artwork Size : 995084 bytes Medium : application/pdf Publisher : NAKALA - <https://nakala.fr> (Huma-Num - CNRS) Version Number : 1.
- BOUCHARD, A. (2024). Sortir de Google Scholar, Scopus ou Web of Science : que valent Lens, Dimensions, OpenAlex et Matilda pour la recherche bibliographique ? [à distance] | URFIST de Paris.
- BOUSSIDAN, A., DUCÉL, F., NÉVÉOL, A. et FORT, K. (2024). What ChatGPT tells us about ourselves.
- BÉNARD, M., MESTIVIER, A., KUBLER, N., ZHU, L., BAWDEN, R., DE LA CLERGERIE, E., ROMARY, L., HUGUIN, M., NOMINÉ, J.-F., PENG, Z. et YVON, F. (2023). MaTOS : Traduction automatique pour la science ouverte. In BOUDIN, F., DAILLE, B., DUFOUR, R., EL, O., HOUBRE, M., JOURDAN, L. et KOOLI, N., éditeurs : *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, pages 8–15, Paris, France. ATALA.
- CUXAC, P. (2024). La fouille de textes en IST : les outils Istex-TDM. Nancy, France.

Références

- CUXAC, P., COLLIGNON, A., GREGORIO, S. et PARMENTIER, F. (2019). Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques. *In 12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances?*
- GREGORIO, S., COLLIGNON, A., PARMENTIER, F. et THOUVENIN, N. (2019). LODEX : des données structurées au web sémantique. *In Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*, Metz, France.
- GUSENBAUER, M. (2022). Search where you will find most : Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, 127(5) :2683–2745.
- HABERT, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? *In BILGER, M., éditeur : Linguistique sur corpus. Etudes et réflexions - Mireille Bilger*, pages 11–58. Presses Universitaires de Perpignan, Perpignan.
- HENDRYCKS, D., BURNS, C., BASART, S., ZOU, A., MAZEIKA, M., SONG, D. et STEINHARDT, J. (2021). Measuring massive multitask language understanding.

Références

- KHAYARI, M., RESZETKO, V., VACHEZ, D., VEDOVOTTO, N., YON, J. et AUBIN, S. (2021). De TermSciences à Loterre : comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR.
- LABADIE, A. et PRINCE, V. (2008). The impact of corpus quality and type on topic based text segmentation evaluation. *In 2008 International Multiconference on Computer Science and Information Technology*, pages 313–319.
- LARIVIÈRE, V., GINGRAS, Y. et ARCHAMBAULT, E. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the American Society for Information Science and Technology*, 60(4) :858–862.
- LIU, Y., CAO, J., LIU, C., DING, K. et JIN, L. (2024). Datasets for Large Language Models : A Comprehensive Survey. arXiv:2402.18041 [cs].
- LÉON, J. (2008). Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages*, 171(3) :12–33. Place : Paris Publisher : Armand Colin.
- MELLET, S. (2002). Corpus et recherches linguistiques : Introduction. *Corpus*, (1).

Références

- RASTIER, F. (2004). Enjeux épistémologiques de la linguistique de corpus. *In* WILLIAM, G., éditeur : *La linguistique de corpus.*, pages 31–46. Presses Universitaires de Rennes, Rennes.
- SINCLAIR, J. (1996). Preliminary recommendations on Corpus Typology. Rapport préliminaire, Expert Advisory Group on Language Engineering Standards.
- STIENNON, N., OUYANG, L., WU, J., ZIEGLER, D. M., LOWE, R., VOSS, C., RADFORD, A., AMODEI, D. et CHRISTIANO, P. (2022). Learning to summarize from human feedback.