



Outil d'analyse de la répartition des langues dans les contenus scientifiques

Samy Assouane

assouane.samy@courrier.uqam.ca
Étudiant à la maîtrise en informatique – IA
UQAM



Chaire de recherche du Québec
Découvrabilité des contenus
scientifiques en français

Plan

1. Introduction

2. Collecte des données

- Types de données
- Sources
- Méthodes de collecte

3. Consolidation des données

4. Visualisation

- Données brutes
- Statistiques générales
- Statistiques par domaines

5. Questions



Introduction

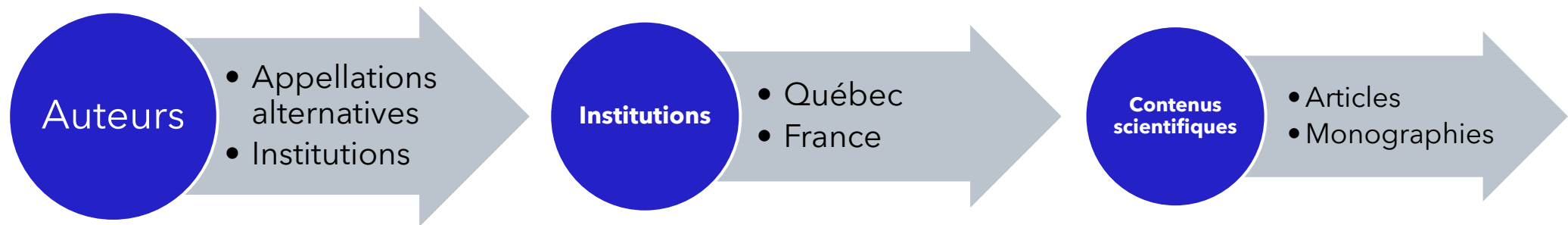
- Plateforme web de récolte et d'analyse de données
- Évolution de la part du français dans les contenus scientifiques
- Auteurs rattachés à une institution québécoise ou française



Collecte des données

Types de données

- Articles et monographies
- Métadonnées : Titre, auteurs, année, langue, domaines, éditeur, institutions, doi, etc.



Collecte des données

Sources

- Plateformes d'indexation



Collecte des données

Méthodes de collecte

- **API (Interface de Programmation d'Application)**
 - Récupérer des données structurées automatiquement
 - Requêtes précises
 - Accès fiable
- **Web scraping (moissonnage)**
 - Extraire les données directement depuis les pages web
 - Accès aux données non disponibles autrement
 - Sensible aux changements de structure HTML



Consolidation des données

- **Analyse des institutions de rattachement des auteurs**
 - Scripts de détection des institutions dans les textes descriptifs des auteurs
 - Mise en relation avec les bases de données d'institutions du Québec et de France
- **Comparaison des données similaires sur les différentes plateformes**
- **Uniformisation des champs**



Visualisation

Données brutes

- Tableaux dynamiques de données avec filtres
- Exportation des données en CSV

 Gestion des Articles

FILTRES 

ARTICLES 

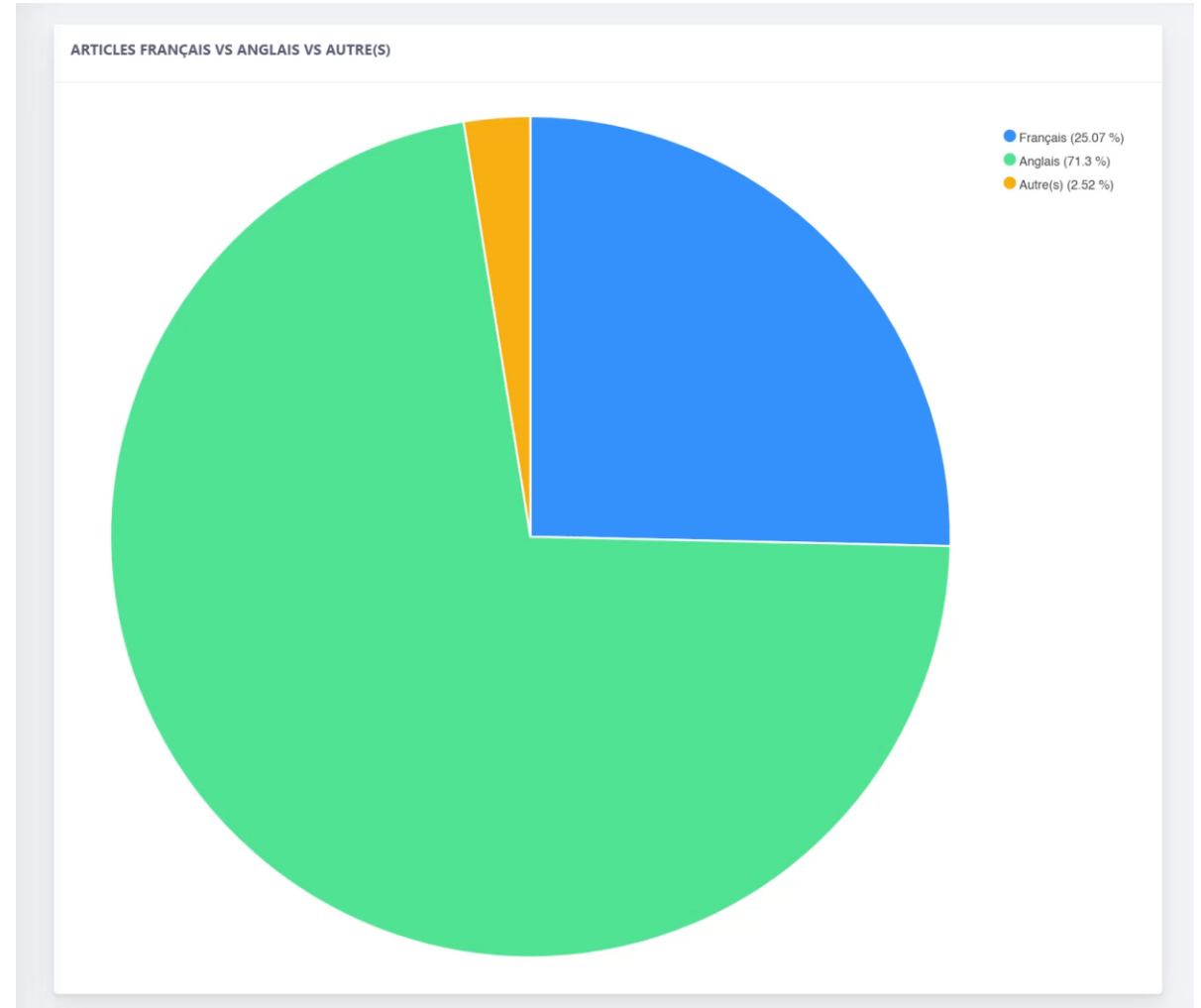
DOI	Article ID	Nom de l'article	Auteurs	Éditeur	Année	Langue	Domaine	Discipline	Institutions	Pays
https://doi.org/10.3322/caac.21492	https://openalex.org/W2889646458	Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries	Freddie Bray Jacques Ferlay Isabelle Soerjomataram Rebecca L. Siegel Lindsey A. Torre Ahmedin Jemal	CA A Cancer Journal for Clinicians	2018	en	Health Sciences	Medicine	Centre International de Recherche sur le Cancer Centre International de Recherche sur le Cancer Centre International de Recherche sur le Cancer American Cancer Society American Cancer Society American Cancer Society	France
https://doi.org/10.3322/caac.21660	https://openalex.org/W3128646645	Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries	Hyuna Sung Jacques Ferlay Rebecca L. Siegel Mathieu Laversanne Isabelle Soerjomataram Ahmedin Jemal Freddie Bray	CA A Cancer Journal for Clinicians	2021	en	Health Sciences	Medicine	American Cancer Society Centre International de Recherche sur le Cancer American Cancer Society Centre International de Recherche sur le Cancer Centre International de Recherche sur le Cancer American Cancer Society Centre International de Recherche sur le Cancer	France



Visualisation

Statistiques générales

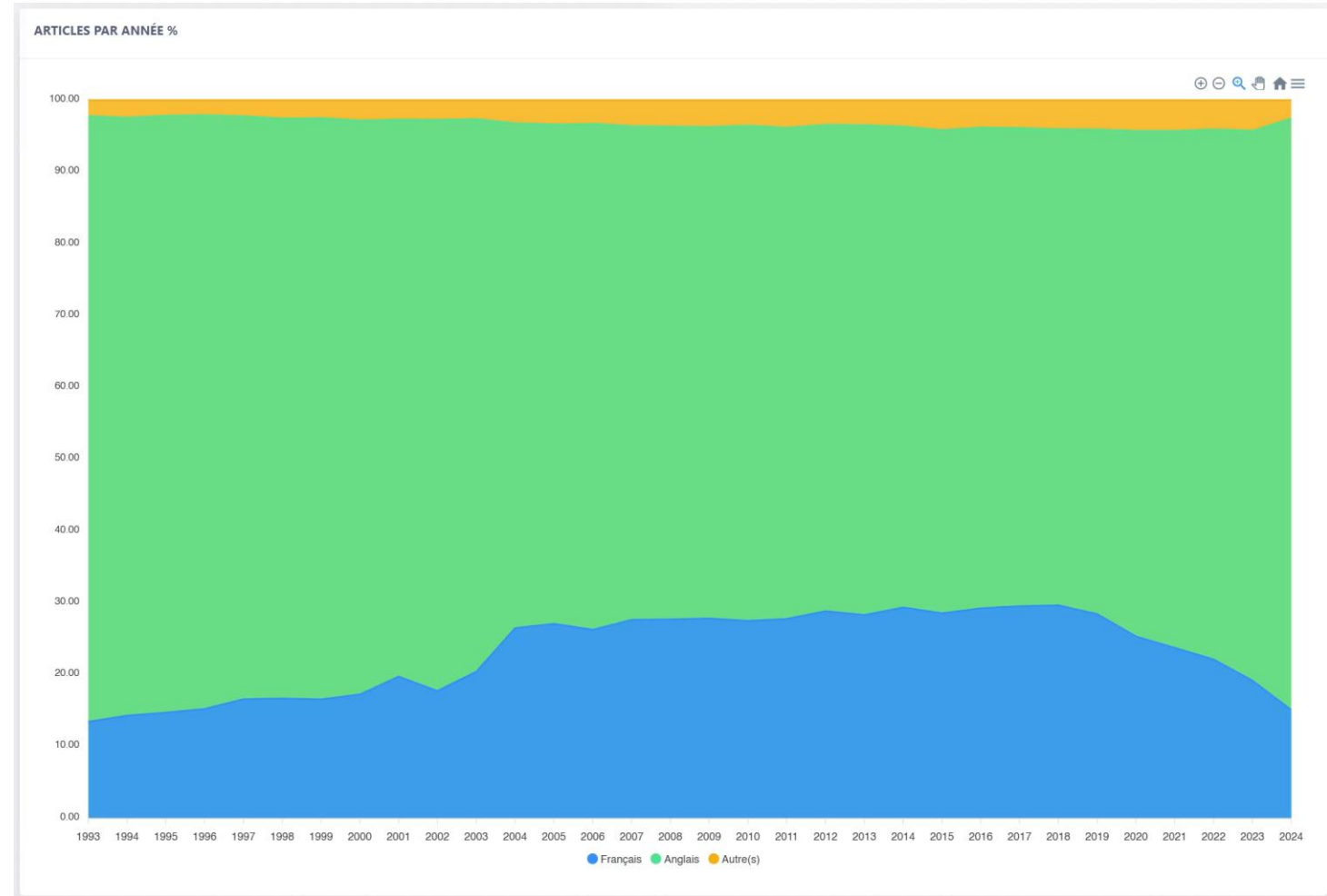
- Graphes dynamiques
- Proportion par langue
- Par type de données (articles ou monographies)
- Par source de données ou global



Visualisation

Statistiques générales

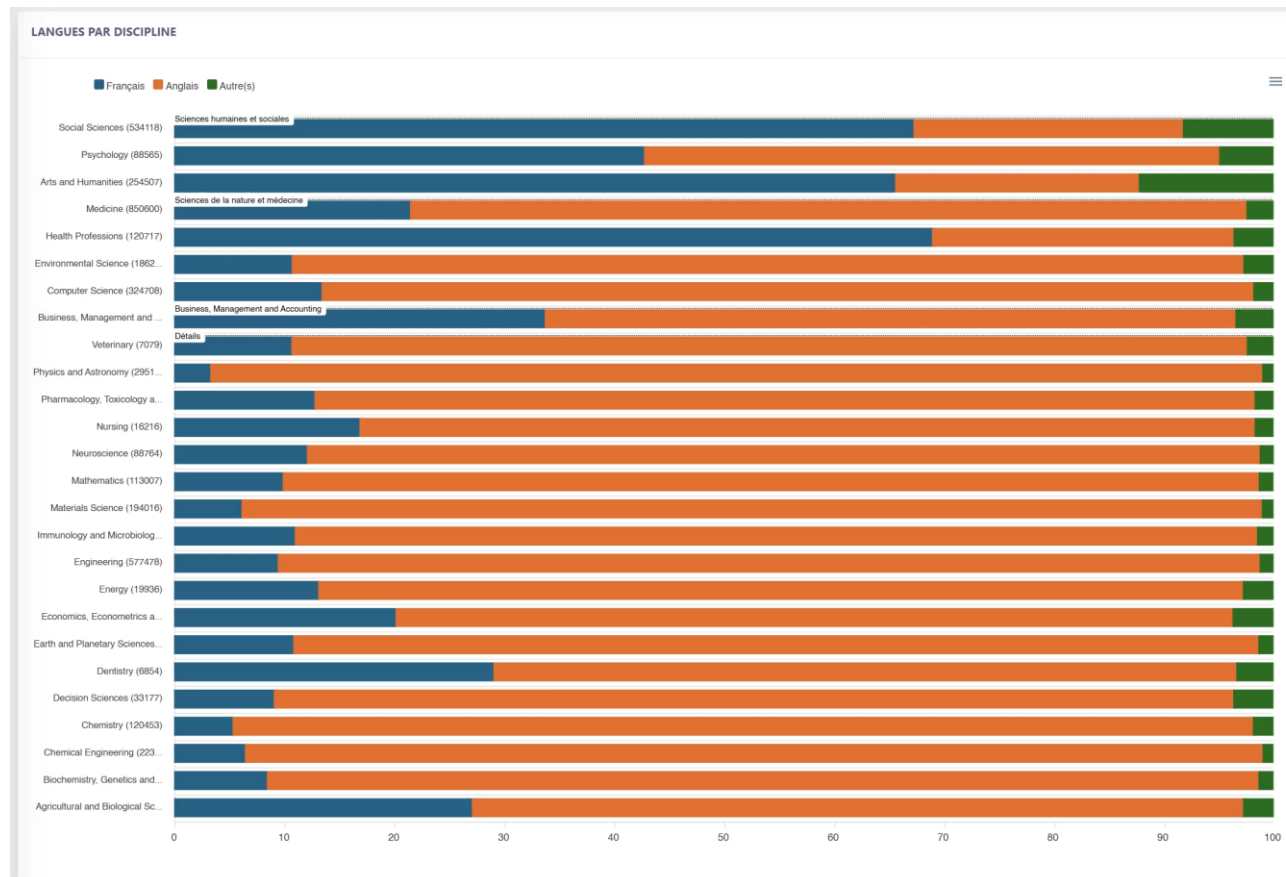
Évolution de la proportion des langues des contenus au cours du temps



Visualisation

Statistiques par domaines

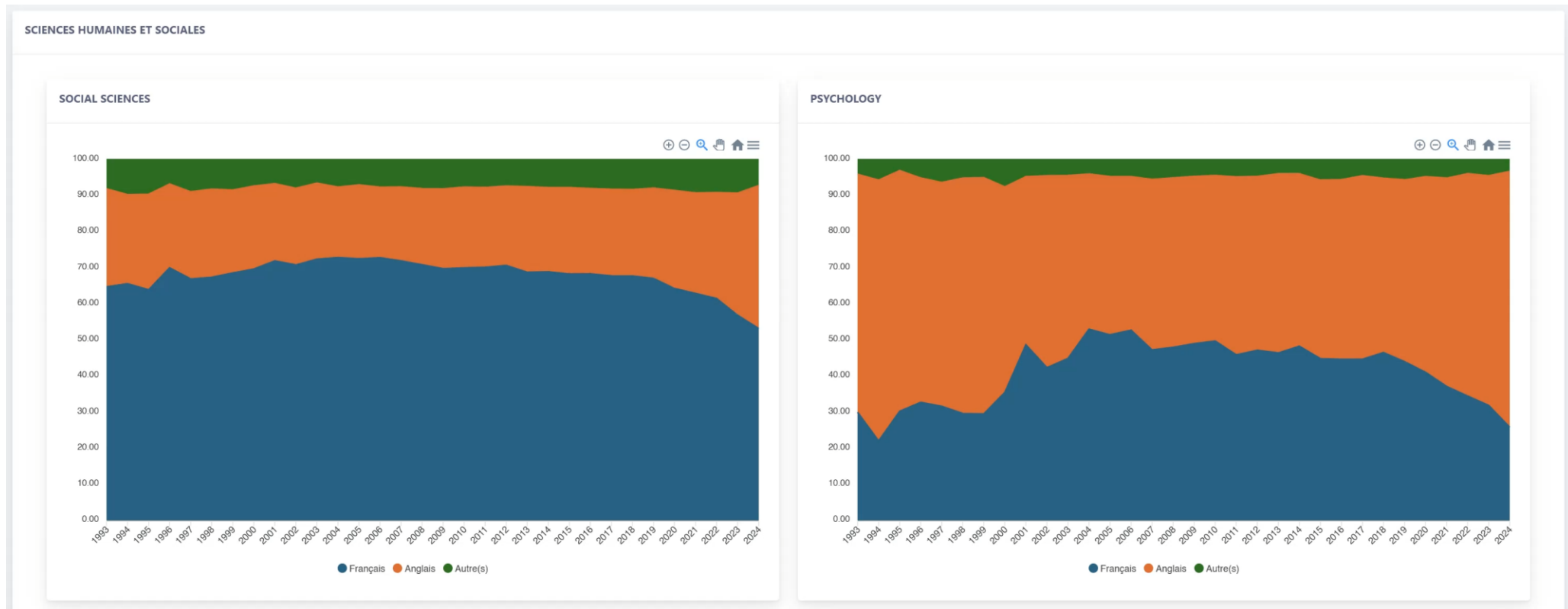
- Répartition des langues des contenus par domaine



Visualisation

Statistiques par domaines

- Répartition des langues des contenus par domaine au cours du temps





Des questions?

Samy Assouane



assouane.samy@courrier.uqam.ca



dcsf.cirst.ca



chaire.dcsf.bsky.social



@chaire-de-recherche-du-quebec-sur-la-decouvrabilite-des-
contenus-scientifiques-en-français

Merci!



Chaire de recherche du Québec
Découvrabilité des contenus
scientifiques en français