



COMPUTER SCIENCE AND STATISTICS
ENGINEER POLYTECH LILLE

Documentation

Data Extraction from Publications and Statistics of
the HAL Database

Anas Nay

School name and address:

POLYTECH Lille
Boulevard Paul Langevin
59655, VILLENEUVE D'ASCQ
CEDEX
03-28-76-73-60

School supervisor: Frédéric
Hoogstoel

Company name and address:

Centre de Recherche en
Informatique, Signal et
Automatique de Lille (CRISTAL)
Université de Lille, Sciences et
technologies, Batiment Esprit,
59655 Villeneuve-d'Ascq

Company supervisor: Mihaly
Petreczky

Academic Year 2024-2025

Contents

1	Contexte et objectifs du projet	4
2	Modes d'utilisation de l'outil	5
2.1	Interface graphique interactive (fichier <code>app.py</code>)	5
2.2	Interface en ligne de commande (fichier <code>main.py</code>)	5
3	Présentation de l'application interactive	6
3.1	Lancement de l'application	6
3.2	Architecture générale	6
3.3	Module d'extraction de données	7
3.3.1	Étape 1 : Extraction des identifiants HAL	7
3.3.2	Vérification et correction des identifiants extraits	9
3.3.3	Étape 2 : Extraction des métadonnées de publications	11
3.4	Module d'analyse de données	14
3.4.1	Génération automatique des visualisations graphiques	14
3.4.2	Consultation interactive du tableau de bord	16
3.4.3	Production de rapports compilés	17
3.4.4	Analyse thématique des mots-clés	18
3.5	Configuration de la sensibilité de correspondance	20
4	Présentation de l'interface ligne de commande	22
4.1	Fonctionnalités principales	22
4.2	Sélection du fichier d'entrée	22
4.3	Options de filtrage	22
4.4	Configuration de la sensibilité	23
4.5	Génération automatique de contenu	23
4.6	Commandes utilitaires	23
4.7	Exemples d'utilisation	23
4.7.1	Extraction simple	23
4.7.2	Extraction avec filtres	23
4.7.3	Extraction avec sensibilité personnalisée	24
4.7.4	Extraction avec génération automatique	24
4.7.5	Commandes d'information	24
4.8	Suivi de progression et résultats	24
4.9	Récapitulatif avant extraction	24
4.10	Organisation des fichiers de sortie	25
5	Module de détection des doublons et homonymes	26
5.1	Principe de fonctionnement	26
5.2	Accès au module depuis l'interface graphique	26
5.3	Configuration et lancement de l'analyse	27
5.3.1	Sélection des fichiers d'entrée	27
5.3.2	Fichier laboratoire optionnel	27
5.4	Interface d'analyse et résultats	27
5.4.1	Onglet Résumé	28
5.4.2	Onglets de résultats détaillés	28

5.5	Types de détection supportés	29
5.6	Traitement automatique des données	29
5.6.1	Options de traitement disponibles	29
5.6.2	Résultats du traitement	29
5.7	Exportation des résultats	30
5.7.1	Fichiers générés	30
5.8	Utilisation en ligne de commande	31
5.8.1	Exemple de session en ligne de commande	31
5.9	Limitations et considérations	32
5.9.1	Dépendance à l'API HAL	32
5.9.2	Cas particuliers	32
5.10	Conclusion	32
6	Annexe	33
6.1	Fichier CSV attendu pour l'extraction	33
6.2	Description du fichier CSV obtenu	33

Note d'avant garde

Cette documentation s'adresse aux utilisateurs possédant des compétences informatiques de base, incluant la capacité d'exécuter des commandes dans un terminal et de gérer l'exécution de fichiers Python. Une familiarité avec les systèmes de gestion de version, notamment GitHub, est également recommandée.

Avant d'utiliser les scripts de ce projet, il est **impératif** de configurer correctement votre environnement de travail. Les scripts, développés en Python, nécessitent l'installation préalable de dépendances spécifiques qui jouent un rôle essentiel dans le traitement des données, la génération de rapports et la visualisation des graphiques.

Installation et configuration

1. Clonage du dépôt

Ouvrez un terminal et exécutez les commandes suivantes :

```
git clone https://github.com/anasnay11/PROJET_HAL_.git
cd PROJET_HAL_
```

2. Création d'un environnement virtuel

Créez et activez un environnement virtuel Python :

```
python -m venv venv
```

Puis activez-le selon votre système d'exploitation :

- **Linux/macOS** : `source venv/bin/activate`
- **Windows** : `venv\Scripts\activate`

3. Installation des dépendances

Une fois l'environnement virtuel activé, installez les packages requis :

```
pip install -r requirements.txt
```

Vérifiez que tous les packages s'installent correctement avant de procéder à l'exécution des scripts.

Points importants à retenir

1. Les fichiers principaux `app.py` et `main.py` se trouvent dans le sous-dossier `python code`. Positionnez-vous dans ce répertoire avant de les exécuter.
2. Le fichier CSV d'entrée doit impérativement contenir les colonnes `'nom'` et `'prenom'` pour être utilisable par l'application.
3. Pour une visualisation optimale des captures d'écran de l'interface, agrandissez votre fenêtre d'application afin de reproduire fidèlement les images présentées dans cette documentation.

1 Contexte et objectifs du projet

Ce projet vise à développer un outil interactif et intuitif pour l'extraction, l'analyse et la visualisation de données scientifiques issues de la plateforme HAL (Hyper Articles en Ligne). HAL constitue l'archive ouverte nationale française qui centralise et diffuse les publications scientifiques produites par les institutions de recherche, laboratoires et chercheurs du territoire.

Face aux volumes croissants de données scientifiques et à la nécessité d'analyser efficacement la production de recherche, cet outil propose une solution complète articulée autour de trois fonctionnalités principales :

- **Extraction automatisée** : récupération ciblée des identifiants d'auteurs et métadonnées de publications selon des critères personnalisables (périodes temporelles, domaines scientifiques, types de documents).
- **Visualisation interactive** : génération automatique de graphiques dynamiques (histogrammes, séries temporelles, diagrammes en barres, nuages de mots) permettant une appréhension immédiate des tendances et patterns dans les données.
- **Reporting professionnel** : production de rapports structurés aux formats PDF et LaTeX, intégrant les visualisations pour une présentation claire et exploitable.

L'architecture de l'outil privilégie l'accessibilité et la simplicité d'usage. Une interface graphique permet à tout utilisateur, indépendamment de ses compétences techniques, d'analyser rapidement les travaux de recherches scientifiques d'un certain laboratoire ou d'un groupe de recherche. Le système requiert uniquement un fichier CSV structuré contenant à minima les noms et prénoms des auteurs d'intérêt.

Le développement de ce système s'est appuyé sur un cas d'usage réel : l'analyse des publications du groupe de recherche MACS¹. Cette approche garantit la pertinence fonctionnelle et la robustesse de l'outil dans des conditions opérationnelles authentiques.

L'ambition finale consiste à délivrer une solution polyvalente et évolutive, adaptée tant aux besoins académiques (évaluations de projets, bilans d'activité scientifique) qu'aux exigences institutionnelles (rapports d'évaluation, tableaux de bord pour directions de laboratoires, aide à la décision stratégique).

¹Le fichier de données de test est disponible en section 5.1

2 Modes d'utilisation de l'outil

L'outil propose deux approches complémentaires pour exploiter les fonctionnalités d'extraction et d'analyse des données scientifiques HAL, chacune adaptée à des profils d'utilisateurs et contextes d'usage distincts.

2.1 Interface graphique interactive (fichier `app.py`)

L'application, développée avec la bibliothèque Tkinter, offre une expérience utilisateur intuitive et accessible. Cette interface graphique structure le processus en deux modules fonctionnels :

- **Module d'extraction** : Configuration et lancement des requêtes de récupération selon des critères personnalisables (fenêtres temporelles, domaines scientifiques, typologies documentaires).
- **Module d'analyse** : Génération de visualisations interactives et export de rapports structurés aux formats PDF et LaTeX.

Cette approche privilégie l'accessibilité et convient particulièrement aux utilisateurs occasionnels ou peu familiers avec les environnements en ligne de commande.

2.2 Interface en ligne de commande (fichier `main.py`)

L'exécution via terminal propose une alternative robuste pour des utilisateurs avancés. Cette méthode permet :

- **Extraction paramétrable** : Définition précise des critères via arguments de ligne de commande, facilitant l'intégration dans des scripts et pipelines de traitement.
- **Génération batch** : Production automatisée de visualisations et rapports sans intervention manuelle, optimisant les traitements sur de gros volumes.
- **Automatisation complète** : Possibilité d'enchaîner extraction, analyse et reporting en une seule commande, idéale pour les analyses récurrentes.

Cette approche maximise l'efficacité et la reproductibilité, particulièrement adaptée aux environnements de recherche nécessitant des analyses systématiques.

Les deux modes garantissent une qualité d'analyse équivalente et accèdent aux mêmes fonctionnalités. Le choix entre interface graphique et ligne de commande dépend principalement des préférences utilisateur, du niveau technique et du contexte d'utilisation.

3 Présentation de l'application interactive

Le fichier `app.py` constitue le cœur de l'interface graphique du projet. Développée avec la bibliothèque `tkinter`, cette application offre une expérience utilisateur intuitive pour l'extraction, l'analyse et la génération de rapports basés sur les données de l'API HAL.

3.1 Lancement de l'application

Deux méthodes permettent d'exécuter l'application :

- **Depuis un IDE** : Exécution directe du fichier Python dans un environnement comme Spyder ou PyCharm
- **Depuis le terminal** : Navigation vers le dossier `PROJET_HAL_/python code` et exécution de la commande `python3 app.py`

3.2 Architecture générale

L'application s'articule autour de trois modules fonctionnels complémentaires :

- **Module Extraction** : Récupération des identifiants HAL et extraction des métadonnées de publications
- **Module Analyse** : Génération de visualisations, analyse des mots-clés et production de rapports
- **Module Détection** : Identification des doublons et homonymes

Au lancement, l'utilisateur accède directement à la section Extraction via l'interface d'accueil :

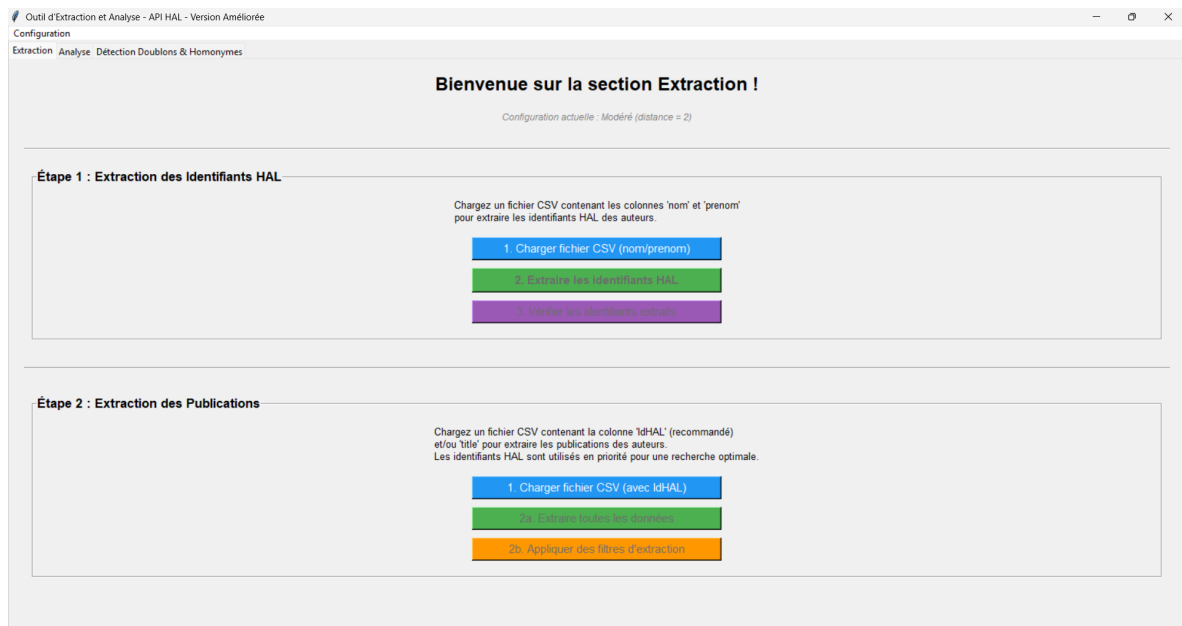


Figure 1: Interface d'accueil - Section Extraction

3.3 Module d'extraction de données

Le processus d'extraction s'effectue en **deux étapes distinctes** pour garantir la précision et la qualité des données extraites. La première étape consiste à identifier les auteurs via leurs identifiants HAL, tandis que la seconde récupère leurs publications.

3.3.1 Étape 1 : Extraction des identifiants HAL

Chargement du fichier source Le processus débute par le chargement d'un fichier CSV contenant la liste des auteurs. Le fichier doit contenir au minimum :

- Soit une colonne **title** avec le nom complet de l'auteur (ex. : "Jean-Luc DUPONT")
- Soit les colonnes **nom** et **prenom** séparément

L'application accepte les deux formats et peut même générer automatiquement les colonnes **nom** et **prenom** à partir de la colonne **title** si elle suit la convention : prénom en casse mixte et nom en MAJUSCULES. Il reste tout de même préférable que le fichier d'entrée contienne les trois colonnes **title**, **nom** et **prenom**.

Une fois le fichier validé, un message de confirmation indique le nombre d'auteurs chargés et les colonnes détectées :

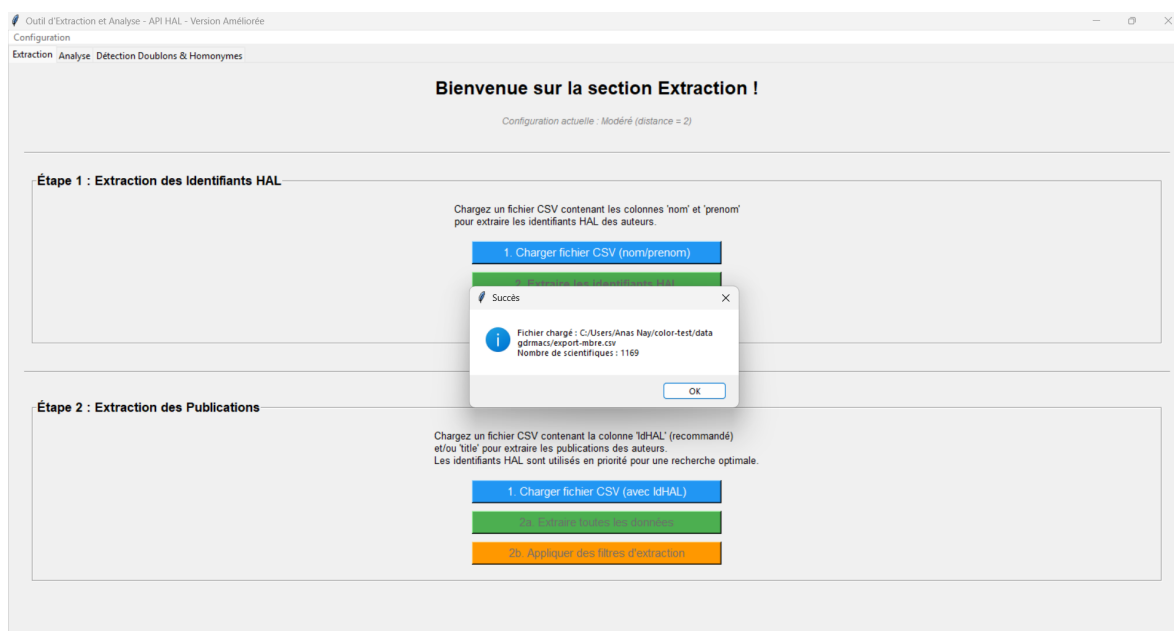


Figure 2: Chargement du fichier CSV (nom/prenom) dans la section Étape 1

Configuration de la sensibilité Avant de lancer l'extraction, il est possible d'ajuster la sensibilité de correspondance des noms via le menu **Configuration > Sensibilité de correspondance**. Cette fonctionnalité, détaillée en section 3.5, permet d'adapter la distance de Levenshtein selon la qualité des données sources.

Lancement de l'extraction des identifiants Un clic sur le bouton "2. Extraire les identifiants HAL" déclenche l'analyse. Une fenêtre récapitulative présente alors les paramètres d'extraction.

Une fois lancée, l'extraction s'exécute avec affichage d'une barre de progression et du compteur d'avancement :

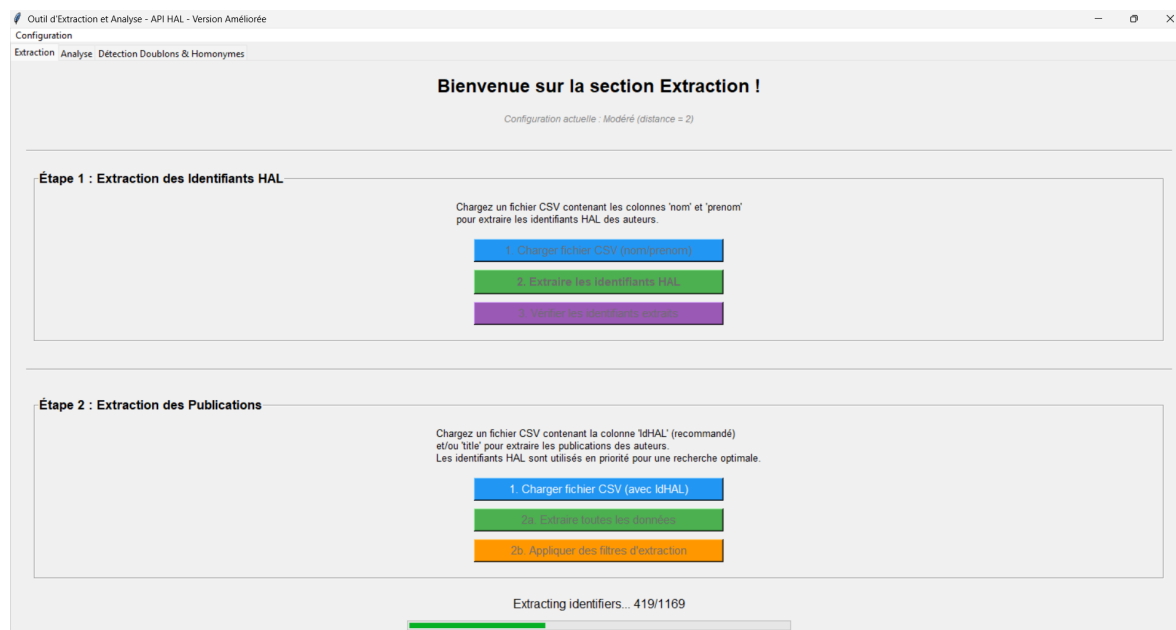


Figure 3: Progression de l'extraction des identifiants des auteurs

À l'issue de l'extraction, un fichier CSV est automatiquement généré dans le dossier **extraction/** avec le nom : **nom_du_fichier_hal_id.csv**. Ce fichier reprend toutes les colonnes du fichier d'origine et ajoute :

- **IdHAL** : L'identifiant HAL de l'auteur (vide si non trouvé)
- **Candidats** : Liste d'identifiants alternatifs séparés par des virgules
- **ID_Atypique** : Indicateur "OUI" ou "NON" selon que l'ID ressemble au nom de l'auteur
- **Details** : Informations techniques de débogage au format JSON

Un message final indique le succès de l'opération et signale les éventuels identifiants atypiques nécessitant une vérification.

3.3.2 Vérification et correction des identifiants extraits

Une fois la première étape terminée, il est possible de vérifier les identifiants extraits (cas d'identifiants atypiques ou candidats multiples). Le bouton "3. Vérifier les identifiants extraits" devient alors accessible. Cette interface de vérification permet une validation manuelle pour les cas nécessitant vérification.

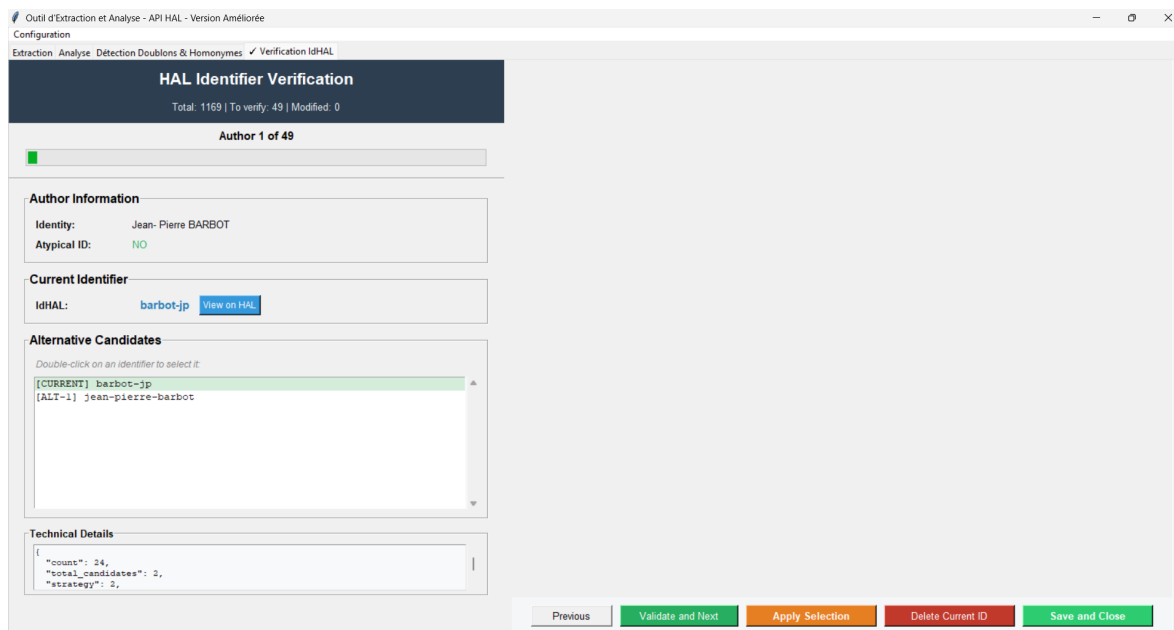


Figure 4: Interface de vérification des identifiants

L'interface de vérification présente pour chaque auteur "problématique" :

- L'identité de l'auteur
- Le statut de l'ID (atypique ou non)
- L'identifiant actuellement assigné
- Un lien menant directement à la liste des publications référencées sous cet identifiant, permettant à l'utilisateur de le vérifier soit-même
- La liste des candidats alternatifs disponibles
- Les détails techniques de l'extraction

Trois actions sont possibles pour chaque auteur :

1. **Valider et passer au suivant** : Confirmer l'identifiant actuel
2. **Sélectionner un candidat alternatif** : Double-clic sur un ID dans la liste ou utilisation du bouton "Appliquer la sélection"
3. **Supprimer l'identifiant** : Bouton "Delete Current ID" si aucun ID ne convient

Un compteur de progression indique l'avancement de la vérification.

À la fin de la vérification (après avoir cliqué sur le bouton "**Sauvegarder et Fermer**"), un nouveau fichier `nom_du_fichier_hal_id_verified.csv` est généré, contenant l'ensemble des auteurs avec les corrections appliquées.

Cas d'usage de la vérification Cette étape de vérification est particulièrement utile pour :

- **Identifiants atypiques** : IDs qui ne ressemblent pas au nom de l'auteur.
- **Homonymes** : Plusieurs auteurs portant le même nom nécessitant une distinction
- **Doublons** : Auteurs présents plusieurs fois dans le fichier source avec des IDs différents
- **Candidats multiples** : Situations où plusieurs IDs plausibles ont été trouvés

La correction des doublons à cette étape permet d'éviter l'extraction redondante de publications lors de l'étape 2, garantissant ainsi la cohérence des analyses ultérieures.

3.3.3 Étape 2 : Extraction des métadonnées de publications

Chargement du fichier avec identifiants Une fois les identifiants HAL extraits (et optionnellement vérifiés), l'étape 2 consiste à récupérer les publications de chaque auteur. Le processus débute par le chargement du fichier CSV contenant les identifiants via le bouton "1. Charger fichier CSV (avec IdHAL)".

Le fichier doit contenir au minimum la colonne `title`. La présence de la colonne `IdHAL` est fortement recommandée. Si cette colonne n'existe pas dans votre fichier, la solution de recours pour extraire les métadonnées de publications sera d'utiliser les noms et prénoms des auteurs (d'où la colonne `title`). L'application utilise en priorité l'identifiant HAL pour les requêtes (méthode la plus précise), puis se rabat sur le nom complet si l'ID est absent ou vide.

Modes d'extraction disponibles Deux approches d'extraction sont proposées selon les besoins analytiques :

Extraction intégrale Ce mode récupère exhaustivement toutes les publications associées aux auteurs listés, sans restriction temporelle, thématique ou typologique. Un clic sur **2a. Extraire toutes les données** ouvre une fenêtre récapitulative présentant le nombre d'auteurs traités et le seuil de sensibilité appliqué :

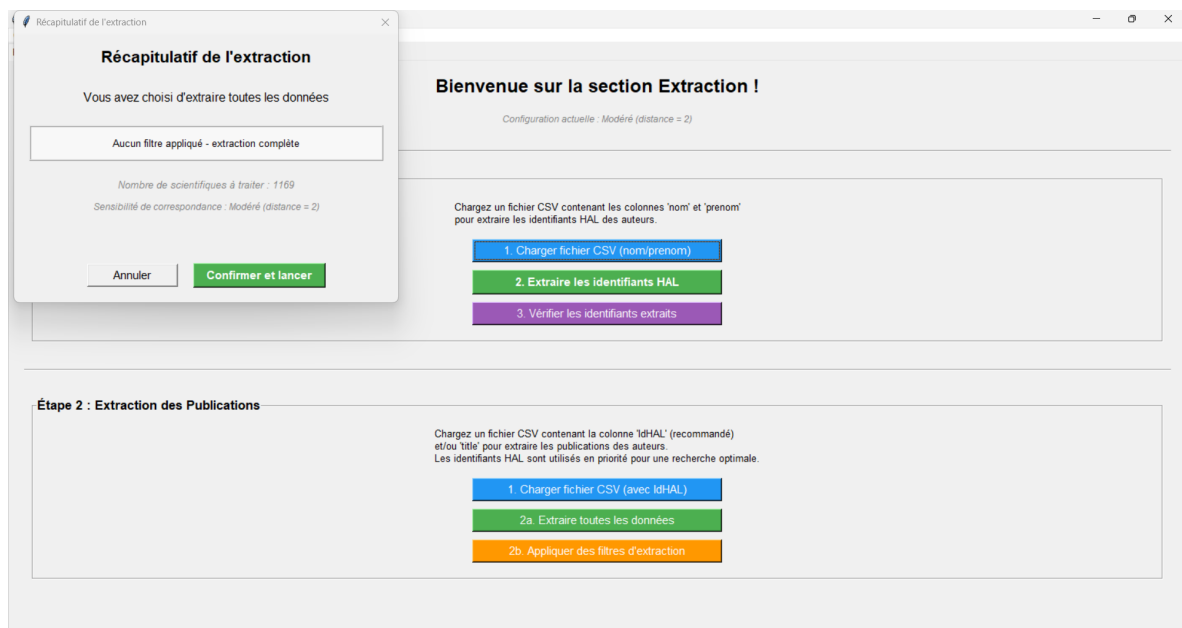


Figure 5: Récapitulatif de l'extraction intégrale

Extraction avec filtres Cette option avancée permet de définir des critères de filtrage multiples via le bouton 2b. **Appliquer des filtres d'extraction :**

- **Fenêtre temporelle :** Restriction à une période spécifique au format AAAA-AAAA (ex. : 2019-2023)
- **Typologie documentaire :** Sélection ciblée (articles, thèses, communications, rapports, etc.)
- **Domaines scientifiques :** Filtrage disciplinaire (informatique, mathématiques, biologie, chimie, etc.)

L'interface de configuration centralise ces paramètres :

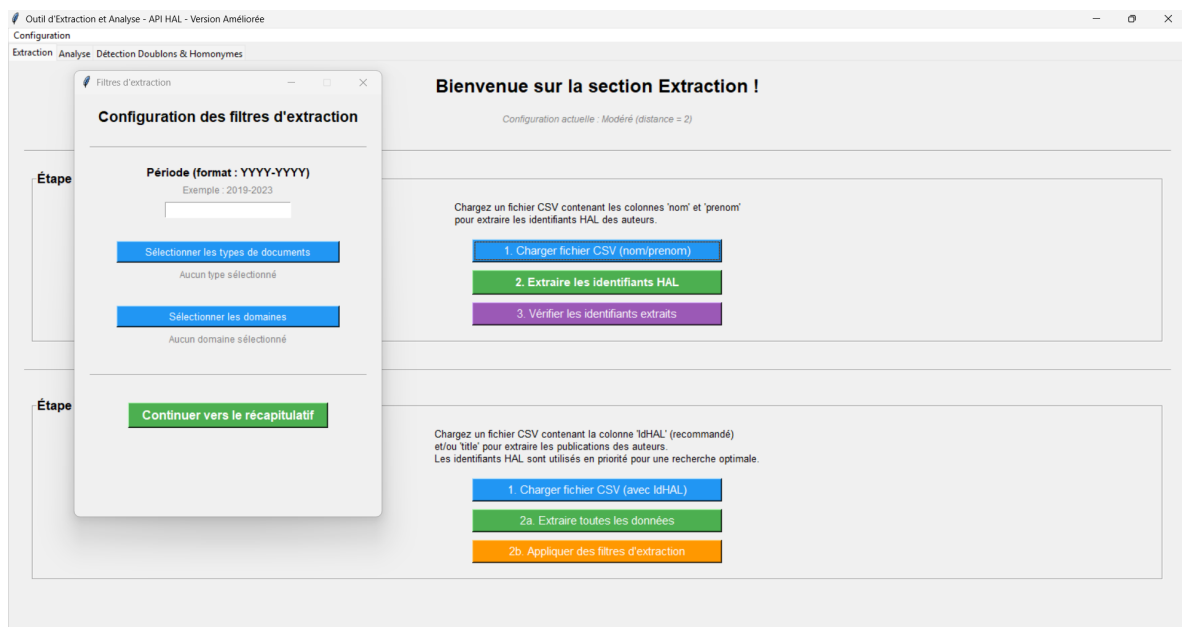


Figure 6: Interface de configuration des filtres

Une fenêtre récapitulative valide la configuration avant lancement.

Exécution et monitoring Le lancement de l'extraction active une barre de progression avec compteur d'avancement.

À l'issue de l'extraction, les résultats sont automatiquement sauvegardés dans le dossier **extraction/** selon une nomenclature standardisée :

- Extraction intégrale : `all_data.csv`
- Extraction filtrée : `all_data-{Domaine}-{Période}-{Type}.csv`

Chaque ligne du fichier généré correspond à une publication et contient les métadonnées suivantes (détaillées en section 6.2) :

- Informations auteur : Nom, Prénom, IdHAL de l'Auteur
- Informations publication : Titre, Docid, Année de Publication, Type de Document, Domaine, Mots-clés, Laboratoire de Recherche affilié
- Collaborations : IdHAL des auteurs de la publication (liste complète)

Un message final confirme le succès de l'opération et indique l'emplacement du fichier généré.

Avantages du processus en deux étapes Cette architecture en deux phases présente plusieurs bénéfices :

- **Traçabilité** : Conservation des identifiants extraits pour vérification ultérieure
- **Réutilisabilité** : Les identifiants validés peuvent servir pour de nouvelles extractions sans re-recherche
- **Précision** : La vérification manuelle élimine les erreurs d'attribution dues aux homonymes
- **Performance** : L'utilisation des IDs HAL accélère considérablement l'extraction des publications
- **Qualité** : Réduction des doublons et des publications erronées dans les analyses finales

3.4 Module d'analyse de données

Le module d'analyse exploite les fichiers CSV générés lors de l'étape 2 de l'extraction (métadonnées de publications) pour produire des visualisations interactives, des analyses thématiques et des rapports structurés.

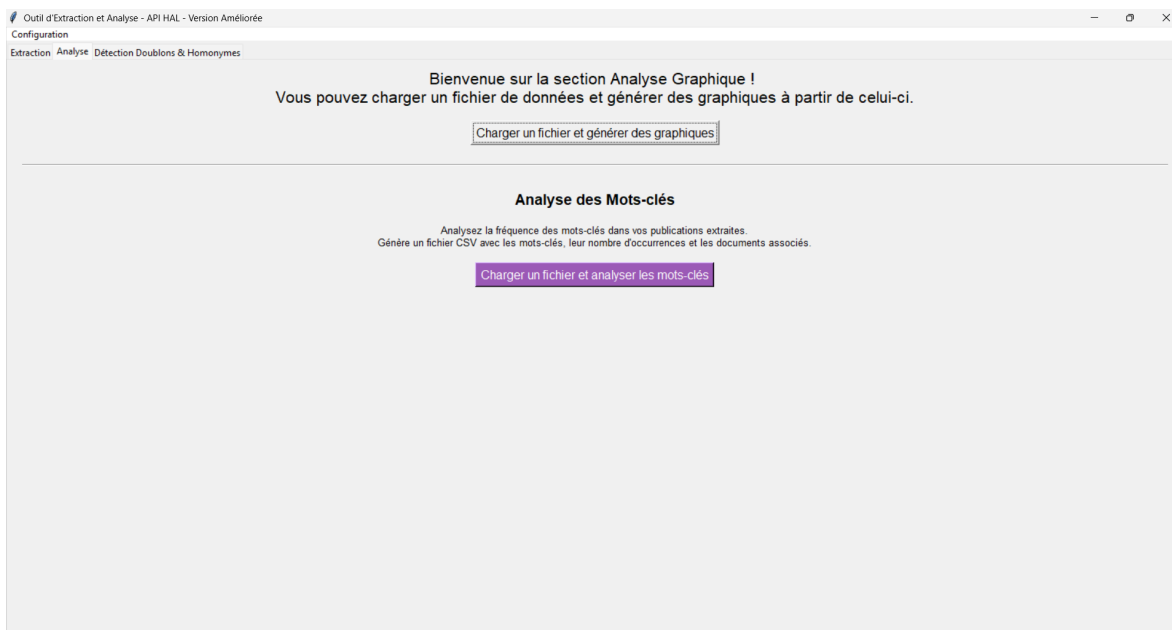


Figure 7: Interface du module d'analyse

3.4.1 Génération automatique des visualisations graphiques

Chargement et préparation des données Le processus débute par le chargement d'un fichier CSV de publications via le bouton **Charger un fichier et générer des graphiques**. Ce fichier doit correspondre à une extraction de publications (fichier `all_data.csv` ou équivalent filtré).

Note importante : Pour garantir la cohérence et la pertinence des analyses visuelles, seules les publications datant de 2005 et après sont prises en compte dans les graphiques. Les données antérieures, souvent incomplètes ou moins représentatives dans HAL, sont automatiquement exclues de la visualisation.

Processus de génération Une fois le fichier chargé, une barre de progression indique l'avancement de la génération des graphiques.

Le système produit automatiquement 9 visualisations distinctes dans deux formats complémentaires :

- **Versions interactives HTML** : Stockées dans le dossier `html/`, développées avec `plotly` pour l'interactivité (zoom, survol, filtrage)
- **Exports statiques PNG** : Stockés dans le dossier `png/`, destinés à l'intégration dans les rapports et présentations

Catalogue des visualisations générées L'application produit les graphiques suivants, chacun apportant un éclairage analytique spécifique :

1. Publications par année

Histogramme interactif présentant l'évolution quantitative de la production scientifique année par année.

2. Distribution des types de documents

Diagramme circulaire affichant la répartition des publications par typologie (articles, thèses, communications, rapports, etc.).

3. Mots-clés les plus fréquents

Diagramme en barres horizontales classant les termes thématiques les plus utilisés dans les publications.

4. Principaux domaines scientifiques

Histogramme identifiant les disciplines les plus représentées dans le corpus.

5. Tendances de publication

Graphique en ligne illustrant l'évolution temporelle de la productivité scientifique avec identification des tendances.

6. Distribution par laboratoire

Diagramme en barres empilées présentant la répartition des publications par structures de recherche.

7. Évolution temporelle par équipe

Graphique multi-courbes suivant l'activité de publication de chaque équipe de recherche dans le temps.

8. Thèses et HDR par année

Diagramme en barres dénombrant les soutenances de thèses et habilitations à diriger des recherches par année.

9. Nuage de mots-clés des thèses

Liste hiérarchique verticale des 15 mots-clés principaux des travaux doctoraux, avec emphase visuelle par taille et couleur.

Un message de confirmation s'affiche à l'issue de la génération, récapitulant le nombre de graphiques créés, le temps d'exécution et l'emplacement des fichiers.

3.4.2 Consultation interactive du tableau de bord

Le bouton **Afficher les graphiques**, activé après génération, lance automatiquement un tableau de bord HTML consolidé dans le navigateur par défaut :



Figure 8: Interface de consultation et export

Ce dashboard centralise l'ensemble des visualisations avec des fonctionnalités interactives avancées permises par la bibliothèque `plotly` :

- Zoom dynamique sur les zones d'intérêt
- Affichage des valeurs exactes au survol
- Filtrage interactif par catégories (clic sur légende)
- Export individuel des graphiques

3.4.3 Production de rapports compilés

Le bouton **Générer un rapport** initie la création d'un document consolidé intégrant l'ensemble des visualisations générées. L'utilisateur sélectionne le format de sortie souhaité :

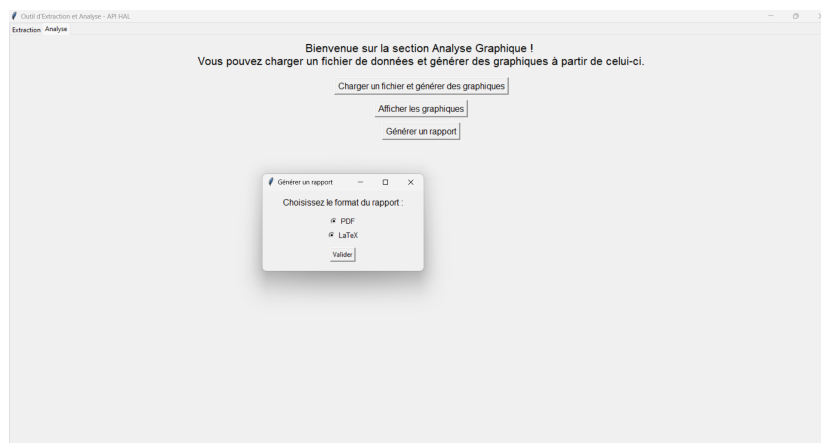


Figure 9: Sélection du format de rapport

Deux formats sont proposés :

- **PDF** : Document directement consultable, idéal pour la diffusion
- **LaTeX** : Code source modifiable pour personnalisation avancée

Les rapports générés sont automatiquement stockés dans le dossier **rapports/** avec les images PNG intégrées pour une consultation offline optimale. Le rapport compile l'ensemble des graphiques dans un document structuré et paginé, accompagné de métadonnées sur l'analyse (période, nombre de publications, etc.).

3.4.4 Analyse thématique des mots-clés

Une fonctionnalité complémentaire permet d'effectuer une analyse approfondie des mots-clés présents dans les publications via le bouton **Charger un fichier et analyser les mots-clés**.

Configuration de l'analyse Le chargement d'un fichier de publications ouvre une fenêtre de configuration permettant de paramétrer l'analyse :

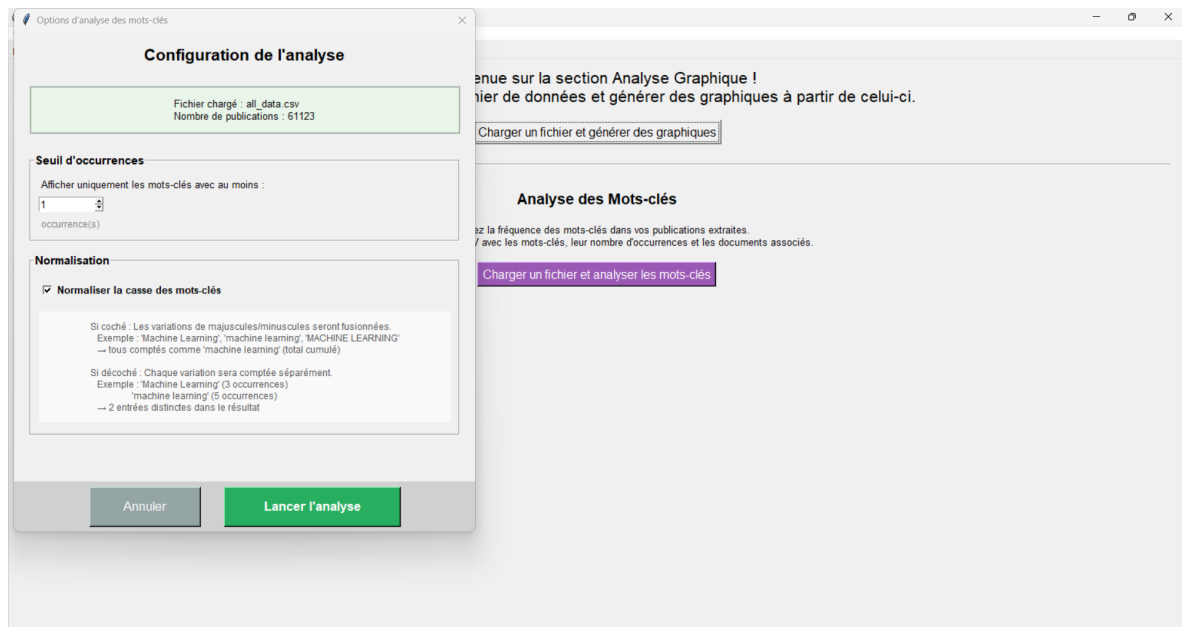


Figure 10: Configuration de l'analyse des mots-clés

Deux paramètres sont configurables :

- **Seuil d'occurrences minimal** : Filtre les mots-clés apparaissant moins de N fois (par défaut : 1)
- **Normalisation de la casse** : Option de fusion des variantes majuscules/minuscules (ex. : "Machine Learning", "machine learning" → "machine learning")

Exécution de l'analyse L'analyse s'exécute avec affichage d'une barre de progression indiquant le nombre de publications traitées.

Fichier de résultats À l'issue de l'analyse, un fichier CSV est automatiquement généré dans le dossier `extraction/` avec la nomenclature : `nom_fichier_keywords_analysis.csv`

Ce fichier structuré contient quatre colonnes :

- **Mot-clé** : Le terme analysé (normalisé si l'option a été activée)
- **Occurrences** : Nombre total d'apparitions du mot-clé dans le corpus
- **Docids** : Identifiants HAL des documents contenant le mot-clé, séparés par des virgules
- **Laboratoires** : Liste unique (sans doublons) des structures de recherche associées aux publications utilisant ce mot-clé, séparées par des virgules

Les résultats sont triés par ordre décroissant d'occurrences, plaçant les mots-clés les plus fréquents en tête de liste.

Un message final confirme le succès de l'analyse et fournit quelques statistiques :

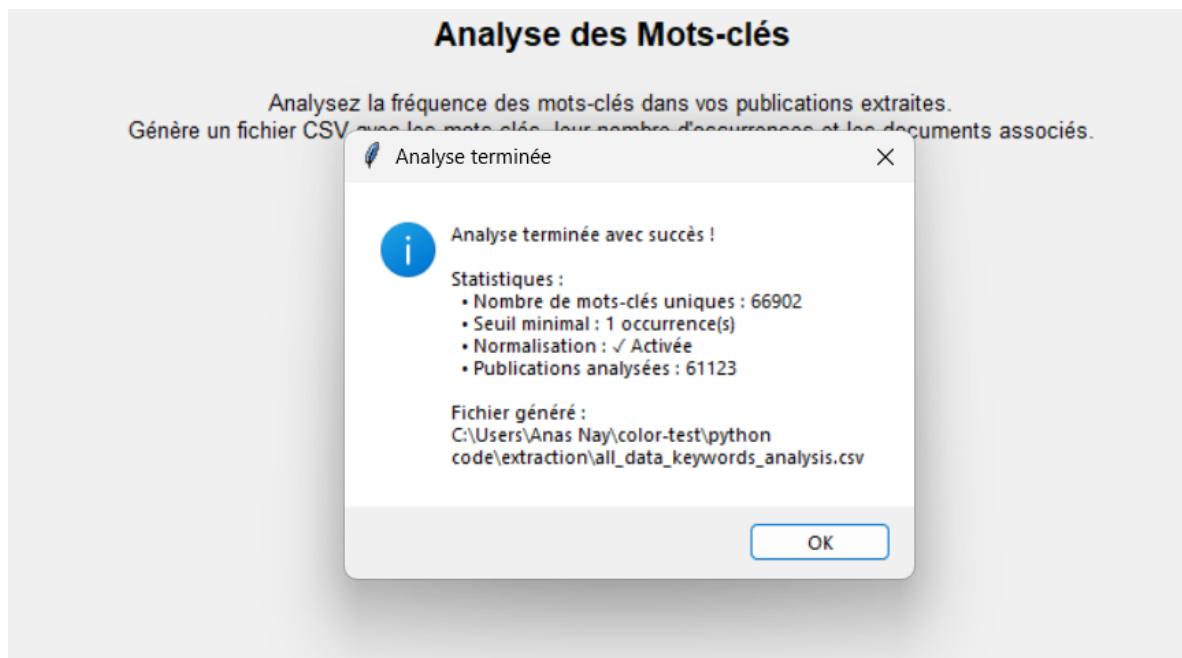


Figure 11: Statistiques Analyse des mots-clés

Applications analytiques Cette analyse thématique permet notamment de :

- Identifier les axes de recherche dominants d'un laboratoire ou d'une équipe
- Cartographier les collaborations inter-laboratoires sur des thématiques communes
- Suivre l'évolution des sujets de recherche en croisant avec les données temporelles
- Détecter les convergences thématiques entre différentes structures

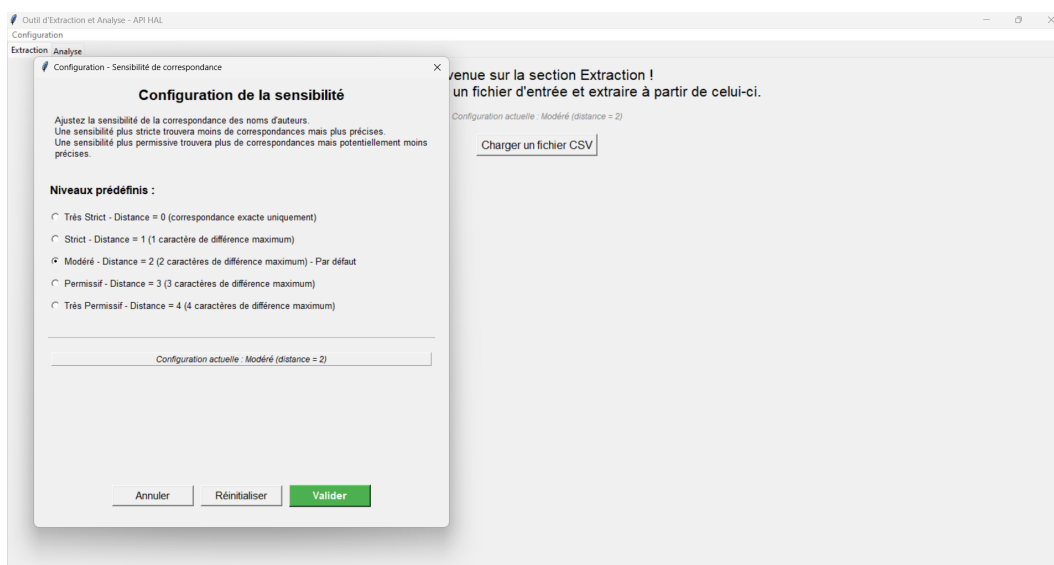
3.5 Configuration de la sensibilité de correspondance

L'application intègre une fonctionnalité avancée permettant de personnaliser le niveau de sensibilité utilisé lors du processus de correspondance des noms et prénoms d'auteurs. Cette configuration est utile pour optimiser la précision de l'extraction en fonction de la qualité et de la cohérence des données d'entrée.

L'accès à ces paramètres s'effectue depuis la barre de menu principale : **Configuration** > **Sensibilité de correspondance**....

La correspondance entre les noms du fichier CSV d'entrée et ceux de la base HAL repose sur la distance de Levenshtein, un algorithme qui mesure la similarité entre deux chaînes de caractères en calculant le nombre minimum d'opérations (insertions, suppressions ou substitutions) nécessaires pour les rendre identiques. Plus cette distance est faible, plus les noms sont similaires. Ce mécanisme gère efficacement les variations d'orthographe, erreurs de frappe, différences d'accentuation ou abréviations courantes.

La fenêtre de configuration présente plusieurs niveaux de sensibilité prédéfinis, chacun correspondant à une valeur de distance spécifique :



Cette interface propose cinq niveaux prédéfinis :

- **Très Strict** (Distance = 0) : Correspondance exacte uniquement, aucune tolérance aux variations
- **Strict** (Distance = 1) : Tolérance d'un seul caractère de différence maximum
- **Modéré** (Distance = 2) : Niveau par défaut, autorise jusqu'à 2 caractères de différence
- **Permissif** (Distance = 3) : Tolérance élargie pour 3 caractères de différence maximum
- **Très Permissif** (Distance = 4) : Niveau le plus tolérant, accepte jusqu'à 4 caractères de différence

Exemple d'application : Pour un nom "Müller" dans le fichier CSV :

- Niveau *Très Strict* : seul "Müller" exact sera trouvé
- Niveau *Modéré* : "Muller", "Müller" seront détectés
- Niveau *Permissif* : inclura également "Miller", "Moller", etc.

Un encadré informatif affiche la configuration actuelle en cours d'utilisation.

Dans l'image précédente, le niveau "Modéré" est sélectionné par défaut avec une distance de 2.

L'interface propose également trois boutons d'action : **Annuler** pour fermer sans sauvegarder, **Réinitialiser** pour revenir aux paramètres par défaut, et **Valider** pour appliquer la nouvelle configuration avant d'extraire les données. Une fois validée, la nouvelle configuration sera confirmée par un message informatif et s'appliquera à toutes les extractions ultérieures.

4 Présentation de l'interface ligne de commande

Le fichier `main.py` constitue une interface en ligne de commande complète pour l'extraction et l'analyse de données scientifiques depuis l'API HAL. Cette approche offre une alternative puissante à l'interface graphique, particulièrement adaptée aux utilisateurs avancés.

4.1 Fonctionnalités principales

Le script propose un ensemble complet d'options permettant de personnaliser finement l'extraction et l'analyse des données :

- **Extraction filtrée** : possibilité de cibler les publications selon des critères spécifiques
- **Gestion de la sensibilité** : configuration du seuil de correspondance des noms d'auteurs
- **Génération automatique** : création de graphiques et rapports en une seule commande
- **Interface interactive** : sélection guidée des fichiers d'entrée

4.2 Sélection du fichier d'entrée

Contrairement aux versions précédentes nécessitant une modification manuelle du chemin de fichier, la version actuelle propose une interface interactive pour sélectionner le fichier CSV contenant les noms des auteurs.

Au lancement du script, une liste numérotée des fichiers CSV disponibles s'affiche :

```
PS C:\Users\Anas Nay\TEST_PROJET_HAL\python code> python .\main.py
Sensitivity threshold used: 2 (moderate)
Available CSV files:
1. export-mbre-test - export-mbre.csv.csv
2. export-mbre.csv

Enter the number of the file you want to use: |
```

L'utilisateur sélectionne simplement le numéro correspondant au fichier souhaité. Cette amélioration rend l'outil plus flexible et élimine la dépendance à la configuration locale.

4.3 Options de filtrage

Le script accepte plusieurs arguments optionnels pour filtrer les résultats d'extraction :

- - - **year** : spécifier une plage d'années (format YYYY-YYYY)
- - - **type** : filtrer par type de document
- - - **domain** : filtrer par domaine scientifique
- - - **threshold** : configurer la sensibilité de correspondance des noms (0-4)

4.4 Configuration de la sensibilité

La nouvelle version intègre la gestion de la sensibilité de correspondance directement en ligne de commande via l'argument `--threshold`. Les niveaux disponibles sont :

- **0** : Très strict (correspondance exacte uniquement)
- **1** : Strict (1 caractère de différence maximum)
- **2** : Modéré (2 caractères de différence maximum) - *Par défaut*
- **3** : Permissif (3 caractères de différence maximum)
- **4** : Très permissif (4 caractères de différence maximum)

4.5 Génération automatique de contenu

Trois nouvelles options permettent la génération automatique de visualisations et rapports :

- `--graphs` : génération automatique de graphiques et ouverture du tableau de bord
- `--reportpdf` : création automatique d'un rapport PDF
- `--reportlatex` : génération automatique d'un rapport LaTeX

4.6 Commandes utilitaires

Le script propose également des commandes d'information :

- `--list-domains` : afficher tous les domaines scientifiques disponibles
- `--list-types` : lister tous les types de documents supportés
- `--list-sensitivity` : détailler les niveaux de sensibilité

4.7 Exemples d'utilisation

4.7.1 Extraction simple

```
python main.py
```

Extraction de toutes les données sans filtres.

4.7.2 Extraction avec filtres

```
python main.py --year 2019--2024 --domain "Mathematics" --type "Theses"
```

Extraction des thèses en mathématiques publiées entre 2019 et 2024.

4.7.3 Extraction avec sensibilité personnalisée

```
python main.py --threshold 1 --domain "Informatique"
```

Extraction avec correspondance stricte des noms pour le domaine informatique.

4.7.4 Extraction avec génération automatique

```
python main.py --graphs --reportpdf --reportlatex
```

Extraction complète avec génération automatique de graphiques et rapports.

4.7.5 Commandes d'information

```
python main.py --list --domains
```

```
python main.py --list-types
```

```
python main.py --list-sensitivity
```

```
python main.py -h
```

4.8 Suivi de progression et résultats

Le script affiche une barre de progression native pendant l'extraction, incluant :

- Pourcentage de completion
- Nombre d'éléments traités
- Estimation du temps restant (ETA)
- Barre visuelle de progression

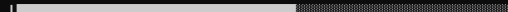
4.9 Récapitulatif avant extraction

Avant de commencer l'extraction, le programme affiche un récapitulatif formaté des paramètres sélectionnés :

```
PS C:\Users\Anas Nay\TEST_PROJET_HAL\python code> python .\main.py
Sensitivity threshold used: 2 (moderate)
Available CSV files:
1. export-mbre-test - export-mbre.csv.csv
2. export-mbre.csv

Enter the number of the file you want to use: 2
You selected the file: export-mbre.csv

=====
EXTRACTION SUMMARY
=====
• Extraction: all data (no filters)
• Matching: sensitivity moderate (distance = 2)
• Outputs: no additional output
=====

Starting extraction...
Extraction in progress:  | 57.0% (666/1169) ETA: 00:13
```

4.10 Organisation des fichiers de sortie

Les résultats sont organisés automatiquement :

- **Fichiers CSV d'extraction** : dossier `extraction/`
- **Graphiques HTML** : dossier `html/`
- **Images PNG** : dossier `png/`
- **Rapports** : dossier `rapports/`

5 Module de détection des doublons et homonymes

Ce module constitue une fonctionnalité avancée de l'outil permettant d'identifier et de traiter automatiquement les publications en double et les cas d'homonymie dans les données extraites de HAL. Cette fonctionnalité s'appuie sur une méthode utilisant les identifiants d'auteur dans la base HAL (`authIdPerson_i`) pour garantir une précision maximale dans la détection.

5.1 Principe de fonctionnement

Le système de détection repose sur un algorithme en plusieurs étapes :

1. **Groupement initial** : Les publications sont regroupées par couple (nom, prénom) d'auteur
2. **Enrichissement via API HAL** : Pour chaque publication, une requête est effectuée vers l'API HAL pour récupérer les métadonnées complètes, notamment les identifiants `authIdPerson_i`
3. **Analyse comparative** : Les publications d'un même auteur sont comparées selon plusieurs critères :
 - Similarité des titres (seuil par défaut : 0.8)
 - Écart temporel entre publications (seuil par défaut : 2 ans)
 - Identifiants HAL officiels
 - Position de l'auteur dans la liste des co-auteurs
4. **Classification automatique** : Les cas détectés sont classifiés en différentes catégories

5.2 Accès au module depuis l'interface graphique

Le module de détection est accessible via l'onglet "Détection Doublons & Homonymes" de l'interface principale :

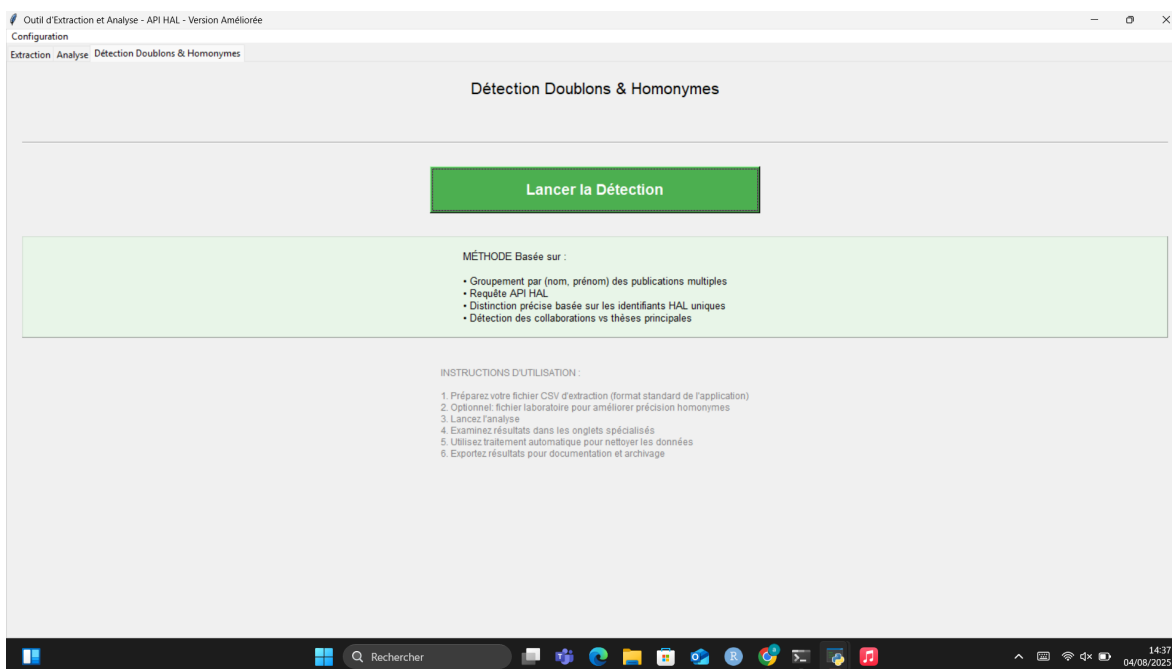


Figure 12: Interface d'accueil du module de détection

L'interface présente les caractéristiques principales de la méthode utilisée et propose un bouton central pour lancer l'analyse. Le processus nécessite un fichier CSV d'extraction préalablement généré par l'outil.

5.3 Configuration et lancement de l'analyse

5.3.1 Sélection des fichiers d'entrée

Le processus d'analyse débute par la sélection du fichier CSV à analyser. Le système propose automatiquement les fichiers du dossier **extraction** et demande si l'utilisateur souhaite utiliser un fichier laboratoire optionnel.

5.3.2 Fichier laboratoire optionnel

L'utilisateur peut optionnellement fournir un fichier contenant les informations de laboratoire des auteurs. Ce fichier doit contenir les colonnes **nom**, **prenom** et **unite_de_recherche**. Cette information supplémentaire améliore significativement la précision de détection des homonymes (en se référant au laboratoire d'affiliation des auteurs, ce qui permet de mieux les différencier).

5.4 Interface d'analyse et résultats

Une fois les fichiers sélectionnés, l'analyse démarre automatiquement dans une interface dédiée organisée en onglets thématiques.

5.4.1 Onglet Résumé

Les résultats sont dans un premier temps présentés dans un onglet résumant l'analyse. En voici un exemple :

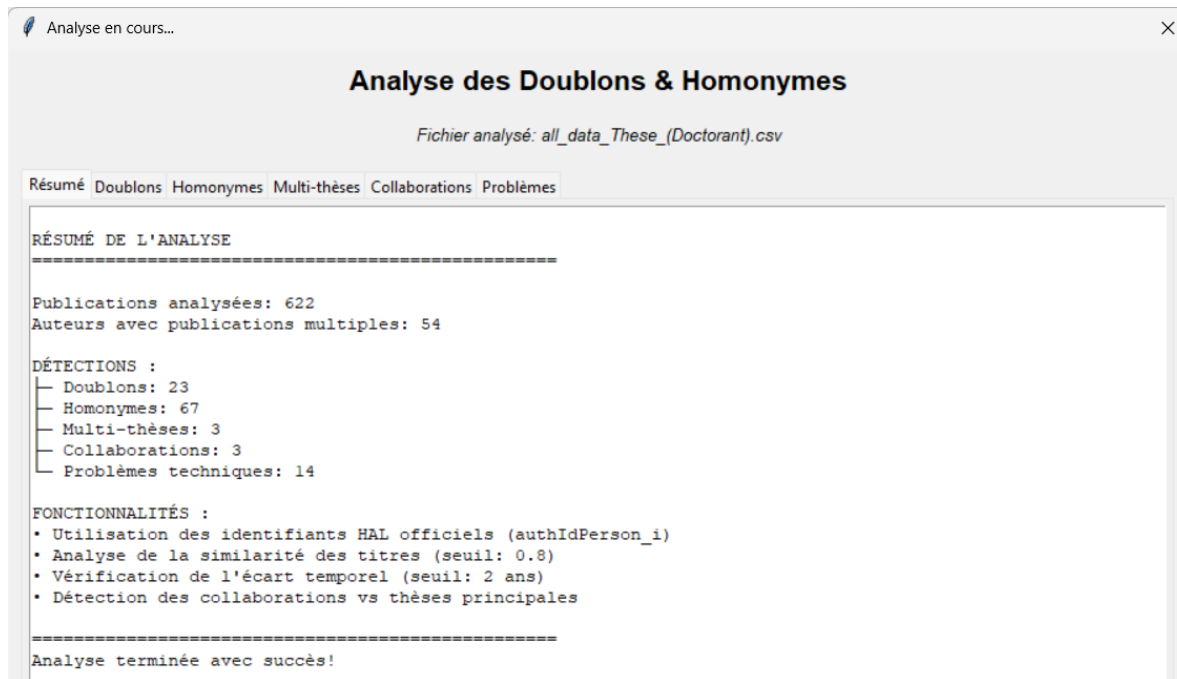


Figure 13: Résumé Détection Doublons et Homonymes

Cet onglet présente une vue d'ensemble des résultats d'analyse avec les statistiques globales incluant :

- Nombre total de publications analysées
- Nombre d'auteurs avec publications multiples
- Détections par catégorie (doublons, homonymes, multi-thèses, collaborations)
- Informations techniques sur la méthode utilisée

5.4.2 Onglets de résultats détaillés

Les différents onglets présentent les cas détectés organisés par catégorie :

- **Onglet Doublons** : Publications identifiées comme doublons potentiels avec scores de similarité, titres comparés et années de publication
- **Onglet Homonymes** : Cas d'homonymie avec critères de différenciation (identifiants HAL différents, domaines scientifiques distincts, laboratoires différents)
- **Onglet Collaborations** : Distinction entre thèses principales et collaborations d'un même auteur
- **Onglet Multi-thèses** : Cas rares d'auteurs avec multiples thèses

- **Onglet Problèmes** : Publications sans `authIdPerson_i` ou métadonnées incomplètes

5.5 Types de détection supportés

Le module identifie plusieurs catégories de problèmes :

Table 1: Types de détection et critères

Type	Critères de détection
Doublons	Même <code>authIdPerson_i</code> , similarité des titres > 0.8 , écart temporel < 2 ans
Homonymes	<code>authIdPerson_i</code> différents, même (nom, prénom), domaines/laboratoires distincts
Multi-thèses	Même <code>authIdPerson_i</code> , écart temporel > 3 ans, faible similarité titres
Collaborations	Position non-principale dans la liste d'auteurs, domaines différents
Problèmes techniques	Absence d' <code>authIdPerson_i</code> , métadonnées HAL incomplètes

5.6 Traitement automatique des données

Le module propose des fonctionnalités de traitement automatique pour nettoyer les données détectées via une interface dédiée accessible après l'analyse.

5.6.1 Options de traitement disponibles

Plusieurs niveaux de traitement sont proposés :

- **Suppression des doublons** : Élimination automatique des publications dupliquées (conservation de la première occurrence)
- **Marquage des homonymes** : Ajout d'une colonne `Homonyme_Potentiel` pour signaler les cas ambigus
- **Suppression des collaborations** : Élimination des collaborations en conservant uniquement les thèses principales
- **Signalement des multi-thèses** : Marquage des cas rares de multiples thèses par un même auteur

5.6.2 Résultats du traitement

Après traitement, le système génère :

- Un fichier CSV nettoyé (suffixe `_nettoye.csv`)
- Un rapport détaillé des actions effectuées
- Des statistiques comparatives avant/après traitement

5.7 Exportation des résultats

Le module offre des fonctionnalités d'exportation complètes pour documenter et archiver les résultats d'analyse. L'utilisateur peut choisir le dossier de destination et le système génère automatiquement plusieurs fichiers spécialisés.

5.7.1 Fichiers générés

L'exportation produit plusieurs fichiers spécialisés :

- `*_doublons_detecte.csv` : Liste détaillée des doublons avec scores de similarité
- `*_homonymes_detecte.csv` : Cas d'homonymie avec informations de différenciation
- `*_multi_theses.csv` : Cas rares de multiples thèses par auteur
- `*_collaborations.csv` : Distinction thèses principales/collaborations
- `*_resume_detecte.txt` : Rapport complet de l'analyse avec méthodologie

5.8 Utilisation en ligne de commande

Le module de détection est également accessible via l'interface en ligne de commande avec l'argument `--analyse` :

```
python main.py --analyse
```

Cette approche propose les mêmes fonctionnalités dans un environnement textuel interactif :

1. Sélection interactive du fichier CSV à analyser
2. Configuration optionnelle du fichier laboratoire
3. Analyse avec affichage des résultats en temps réel
4. Options de traitement et d'exportation intégrées

5.8.1 Exemple de session en ligne de commande

```
# Lancement de l'analyse
python main.py --analyse

# Selection du fichier (interface interactive)
Fichiers disponibles dans 'extraction/':
1. all_data_These_Doctorant.csv
2. publications_2020-2024.csv
Selectionnez un fichier (1-2): 1

# Configuration du fichier laboratoire
Utiliser un fichier laboratoire ? (o/n): o
Fichier laboratoire selectionne: laboratoires.csv

# Analyse en cours avec barre de progression
Analyse en cours... (patience requise - interrogation API HAL)
Analyse de Dupont Jean (3 publications) - 15/120...

# Affichage des resultats details
```

```
RESULTATS DE L'ANALYSE
```

```
Publications-analysees:-450
Auteurs-avec-publications-multiples:-120
Doublons-detectes:-23
Homonymes-detectes:-8
Collaborations-detectees:-15
```


5.9 Limitations et considérations

5.9.1 Dépendance à l'API HAL

La précision de la détection dépend de la qualité des métadonnées HAL :

- Publications sans `authIdPerson_i` : traitement par similarité de titres uniquement
- Métadonnées incomplètes : classification dans les "problèmes techniques"
- Évolution des données HAL : possibles variations dans le temps

5.9.2 Cas particuliers

Certains cas nécessitent une validation manuelle :

- Changements de nom d'auteur (mariage, etc.)
- Collaborations internationales avec variations d'orthographe
- Publications avec erreurs de métadonnées côté HAL

5.10 Conclusion

Le module de détection des doublons et homonymes constitue un outil puissant pour améliorer la qualité des données d'analyse bibliométrique. En combinant l'utilisation des identifiants officiels HAL avec des heuristiques avancées, il permet une détection automatique fiable de la majorité des cas problématiques tout en proposant des solutions de traitement adaptées.

L'intégration de ce module dans l'outil global garantit une chaîne de traitement complète, depuis l'extraction initiale jusqu'à l'analyse finale, en passant par une phase de nettoyage et de validation des données qui améliore significativement la pertinence des résultats obtenus, et donc des statistiques générées par la suite.

6 Annexe

6.1 Fichier CSV attendu pour l'extraction

Pour télécharger et visualiser le type de fichier CSV requis pour l'extraction, veuillez utiliser le lien suivant:

Téléchargez le fichier CSV ici

6.2 Description du fichier CSV obtenu

Le fichier CSV obtenu après l'extraction des données, que ce soit via l'interface de l'application ou en exécutant le fichier `main.py`, présente une structure uniforme. Le fichier contiendra les mêmes informations peu importe la méthode utilisée pour l'extraction. Chaque ligne du fichier CSV représente une publication. Voici une description des colonnes typiques que l'on trouve dans ce fichier :

Table 2: Description des colonnes du fichier CSV

Colonne	Description
Nom	Le nom de famille de l'auteur.
Prénom	Le prénom de l'auteur.
IdHAL de l'Auteur	Identifiant HAL de l'auteur.
IdHAL des auteurs de la publication	Identifiant HAL des auteurs de la publication.
Titre	Le titre de la publication.
Docid	L'identifiant de la publication.
Année de Publication	L'année de publication.
Type de Document	Le type de document, par exemple article, thèse, etc.
Domaine	Le domaine scientifique de la publication.
Mots-clés	Les mots-clés associés à la publication.
Laboratoire de Recherche	Le laboratoire ou le centre de recherche associé à l'auteur de la publication.