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

Foreword

This documentation is intended for users with basic computer skills, including the ability to execute commands in a terminal and manage Python file execution. Familiarity with version control systems, particularly GitHub, is also recommended.

Before using the scripts in this project, it is **imperative** to properly configure your working environment. The scripts, developed in Python, require the prior installation of specific dependencies that play an essential role in data processing, report generation, and graph visualization.

Installation and Configuration

1. Repository Cloning

Open a terminal and execute the following commands:

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

2. Virtual Environment Creation

Create and activate a Python virtual environment:

```
python -m venv venv
```

Then activate it according to your operating system:

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

3. Dependencies Installation

Once the virtual environment is activated, install the required packages:

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

Verify that all packages install correctly before proceeding to script execution.

Important Points to Remember

1. The main files `app.py` and `main.py` are located in the `python code` subfolder. Navigate to this directory before executing them.
2. The input CSV file must contain the columns `'nom'` and `'prenom'` to be usable by the application.
3. For optimal visualization of the interface screenshots, maximize your application window to faithfully reproduce the images presented in this documentation.

1 Project Context and Objectives

This project aims to develop an interactive and intuitive tool for the extraction, analysis, and visualization of scientific data from the HAL platform (Hyper Articles en Ligne). HAL constitutes the French national open archive that centralizes and disseminates scientific publications produced by research institutions, laboratories, and researchers across the territory.

Faced with the growing volumes of scientific data and the need to efficiently analyze research output, this tool offers a comprehensive solution built around three main functionalities:

- **Automated extraction:** targeted retrieval of author identifiers and publication metadata according to customizable criteria (time periods, scientific domains, document types).
- **Interactive visualization:** automatic generation of dynamic graphics (histograms, time series, bar charts, word clouds) enabling immediate understanding of trends and patterns in the data.
- **Professional reporting:** production of structured reports in PDF and LaTeX formats, integrating visualizations for clear and actionable presentation.

The tool's architecture prioritizes accessibility and ease of use. A graphical interface allows any user, regardless of their technical skills, to quickly analyze the scientific work of a particular laboratory or research group. The system only requires a structured CSV file containing at minimum the names and first names of the authors of interest.

The development of this system was based on a real use case: the analysis of publications from the MACS research group[1]. This approach ensures functional relevance and tool robustness under authentic operational conditions.

The final ambition is to deliver a versatile and scalable solution, suited both to academic needs (project evaluations, scientific activity reports) and institutional requirements (evaluation reports, dashboards for laboratory management, strategic decision support).

[1]The test data file is available in section 5.1

2 Tool Usage Modes

The tool offers two complementary approaches to exploit HAL scientific data extraction and analysis functionalities, each adapted to distinct user profiles and usage contexts.

2.1 Interactive Graphical Interface (`app.py` file)

The application, developed with the Tkinter library, provides an intuitive and accessible user experience. This graphical interface structures the process into two functional modules:

- **Extraction module:** Configuration and launching of retrieval queries according to customizable criteria (time windows, scientific domains, document typologies).
- **Analysis module:** Generation of interactive visualizations and export of structured reports in PDF and LaTeX formats.

This approach prioritizes accessibility and is particularly suitable for occasional users or those unfamiliar with command-line environments.

2.2 Command-Line Interface (`main.py` file)

Terminal execution offers a robust alternative for advanced users. This method enables:

- **Configurable extraction:** Precise definition of criteria via command-line arguments, facilitating integration into scripts and processing pipelines.
- **Batch generation:** Automated production of visualizations and reports without manual intervention, optimizing processing of large volumes.
- **Complete automation:** Ability to chain extraction, analysis, and reporting in a single command, ideal for recurring analyses.

This approach maximizes efficiency and reproducibility, particularly suited to research environments requiring systematic analyses.

Both modes guarantee equivalent analysis quality and access the same functionalities. The choice between graphical interface and command line depends primarily on user preferences, technical level, and usage context.

3 Interactive Application Overview

The `app.py` file constitutes the core of the project's graphical interface. Developed with the `tkinter` library, this application offers an intuitive user experience for extraction, analysis, and report generation based on HAL API data.

3.1 Application Launch

Two methods allow executing the application:

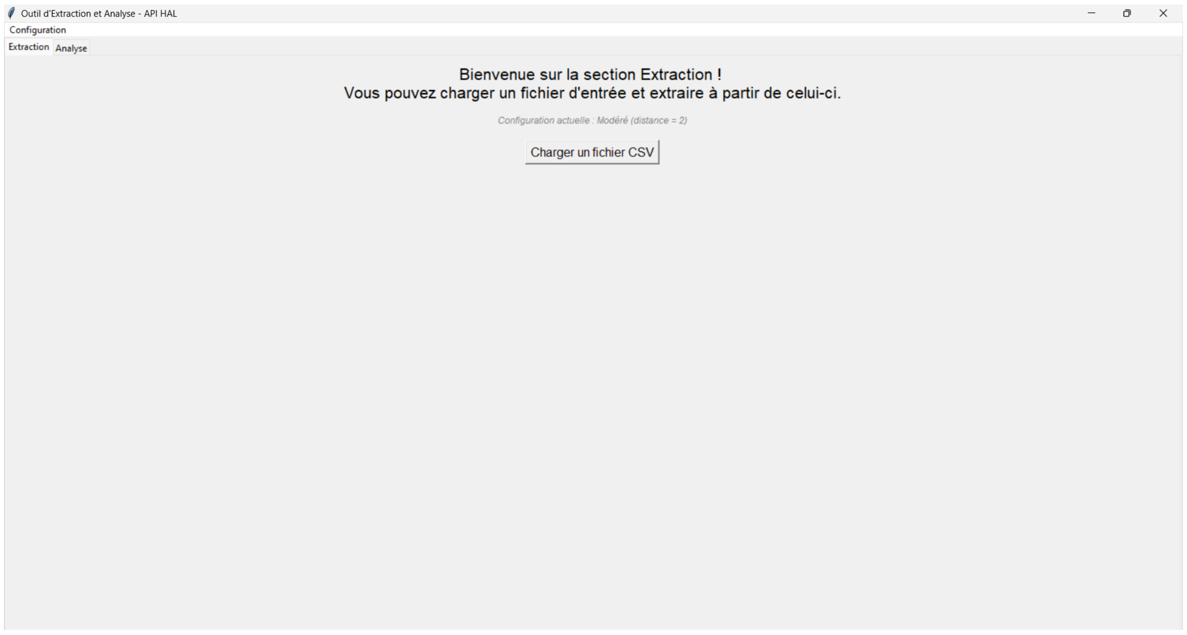
- **From an IDE:** Direct execution of the Python file in an environment such as Spyder or PyCharm
- **From the terminal:** Navigation to the `PROJET_HAL_/python code` folder and execution of the command `python3 app.py`

3.2 General Architecture

The application is built around two complementary functional modules:

- **Extraction Module:** Retrieval and filtering of scientific data from HAL
- **Analysis Module:** Generation of visualizations and production of reports

Upon launch, the user directly accesses the Extraction section via the welcome interface:



Bienvenue sur la section Extraction !
Vous pouvez charger un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

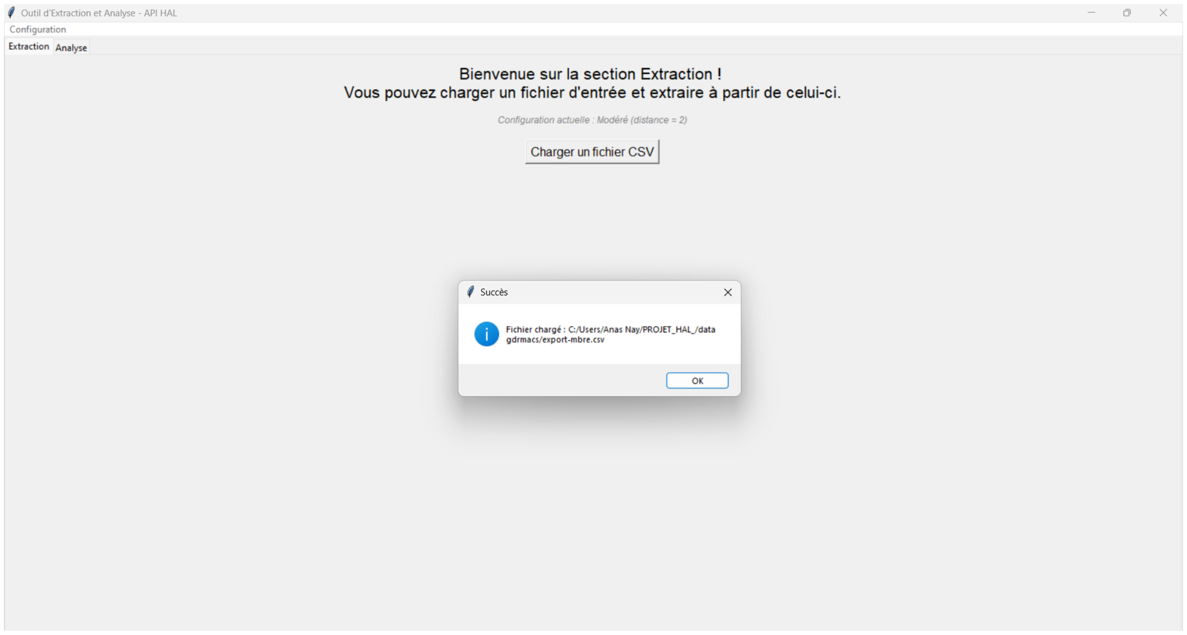
Figure 1: Welcome Interface - Extraction Section

3.3 Data Extraction Module

The extraction process is initialized by loading a structured CSV file containing the identities of target authors (mandatory `nom` and `prenom` columns). The extracted data enriches this initial file with author metadata (HAL identifiers, affiliations) and detailed information about their publications (titles, types, domains, etc.). A comprehensive description of these fields is available in section 5.2.

3.3.1 Source Data Loading

The interface provides a central button to select the input CSV file. Once the file is validated, a confirmation message is displayed:



Bienvenue sur la section Extraction !
Vous pouvez charger un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

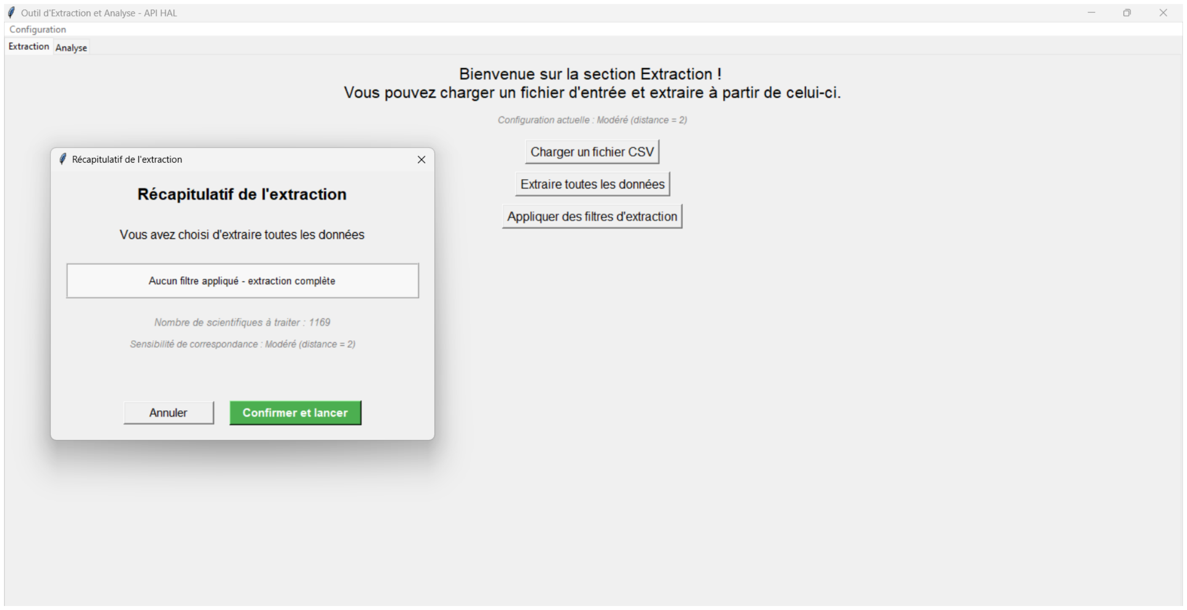
Fichier chargé : C:/Users/Anas Nay/PROJET_HAL_/data gdrmacs/export-mbre.csv

Figure 2: Successful Loading Confirmation

3.3.2 Available Extraction Modes

Two extraction approaches are offered according to analytical needs:

Complete Extraction This mode exhaustively retrieves all publications associated with the listed authors, without temporal, thematic, or typological restrictions. A summary window presents the number of authors processed and the sensitivity threshold applied (detailed in section 3.5):


Outil d'Extraction et Analyse - API HAL

Configuration

Extraction Analyse

Bienvenue sur la section Extraction !
Vous pouvez charger un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

Charger un fichier CSV
Extraire toutes les données
Appliquer des filtres d'extraction

Récapitulatif de l'extraction

Récapitulatif de l'extraction

Vous avez choisi d'extraire toutes les données

Aucun filtre appliqué - extraction complète

Nombre de scientifiques à traiter : 1169

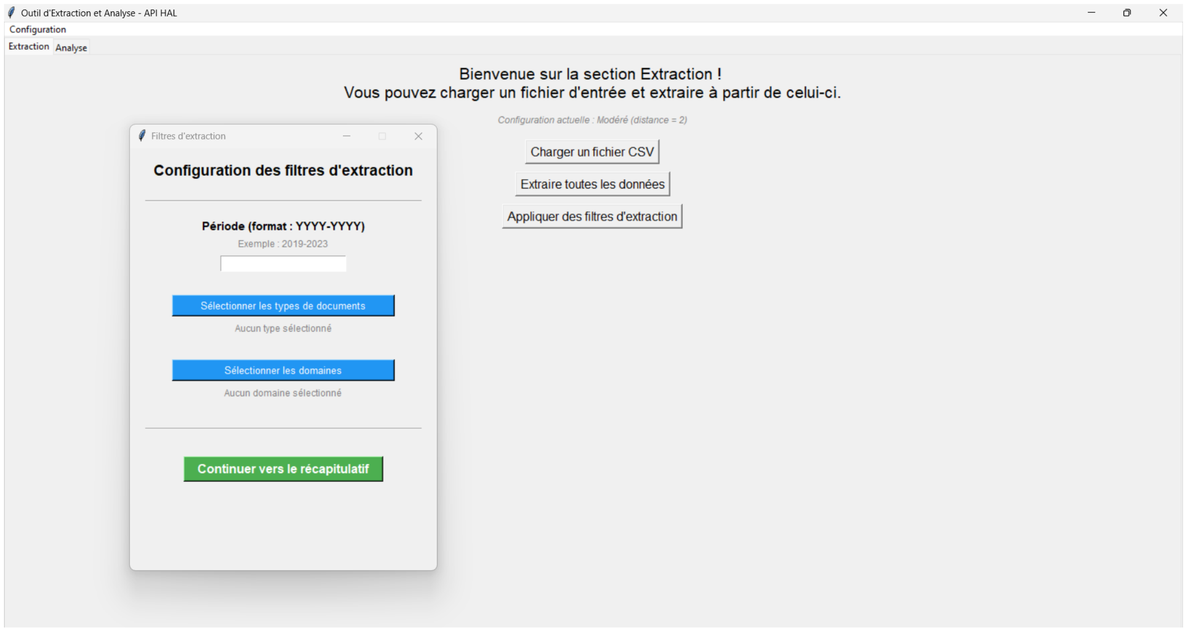
Sensibilité de correspondance : Modéré (distance = 2)

Figure 3: Complete Extraction Summary

Filtered Extraction This advanced option allows defining multiple filtering criteria:

- **Time window:** Restriction to a specific period (e.g., 2019-2022)
- **Document typology:** Targeted selection (articles, theses, communications, reports)
- **Scientific domains:** Disciplinary filtering (computer science, mathematics, biology, etc.)

The configuration interface centralizes these parameters:


Outil d'Extraction et Analyse - API HAL

Configuration

Extraction Analyse

Bienvenue sur la section Extraction !
Vous pouvez charger un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

Charger un fichier CSV
Extraire toutes les données
Appliquer des filtres d'extraction

Filtres d'extraction

Configuration des filtres d'extraction

Période (format : YYYY-YYYY)

Exemple : 2019-2023

Sélectionner les types de documents

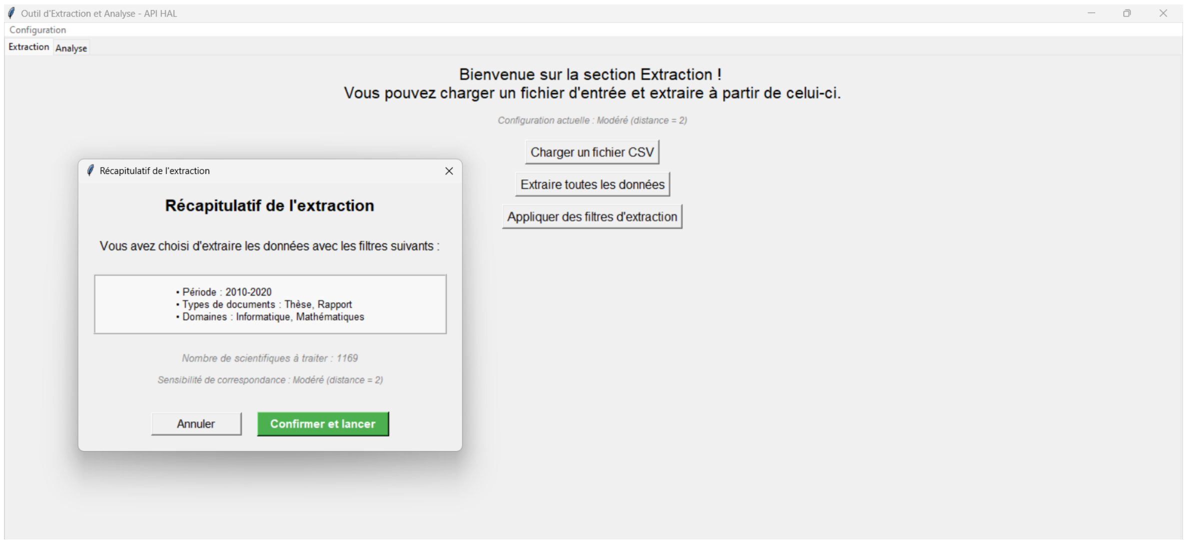
Aucun type sélectionné

Sélectionner les domaines

Aucun domaine sélectionné

Figure 4: Filter Configuration Interface

A summary window validates the configuration before launch. The example below illustrates filtering for theses and reports in computer science and mathematics for the period 2010-2020:


Outil d'Extraction et Analyse - API HAL
Configuration
Extraction
Analyse

Bienvenue sur la section Extraction !
Vous pouvez charger un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

Charger un fichier CSV
Extraire toutes les données
Appliquer des filtres d'extraction
Récapitulatif de l'extraction

Récapitulatif de l'extraction

Vous avez choisi d'extraire les données avec les filtres suivants :

• Période : 2010-2020
• Types de documents : Thèse, Rapport
• Domaines : Informatique, Mathématiques

Nombre de scientifiques à traiter : 1169

Sensibilité de correspondance : Modéré (distance = 2)

Figure 5: Extraction Summary with Applied Filters

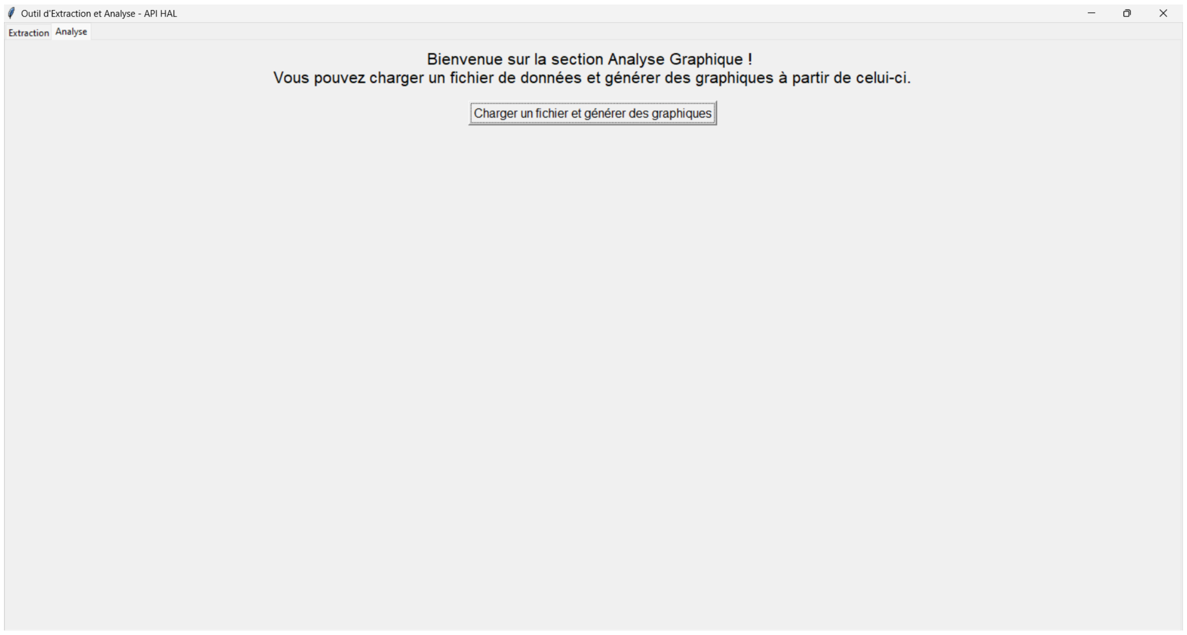
3.3.3 Execution and Monitoring

Launching the extraction activates a progress bar with advancement percentage. An emergency stop button allows process interruption if necessary.
Upon completion of the extraction, results are automatically saved in the `extraction` folder according to standardized nomenclature:

- Complete extraction: `all_data.csv`
- Filtered extraction: `all_data_{Domain}_{Period}_{Type}.csv`

3.4 Data Analysis Module

The analysis module exploits CSV files generated during extraction to produce interactive visualizations and structured reports.


Outil d'Extraction et Analyse - API HAL
Extraction
Analyse
Charger un fichier et générer des graphiques

Bienvenue sur la section Analyse Graphique !
Vous pouvez charger un fichier de données et générer des graphiques à partir de celui-ci.

Figure 6: Analysis Module Interface

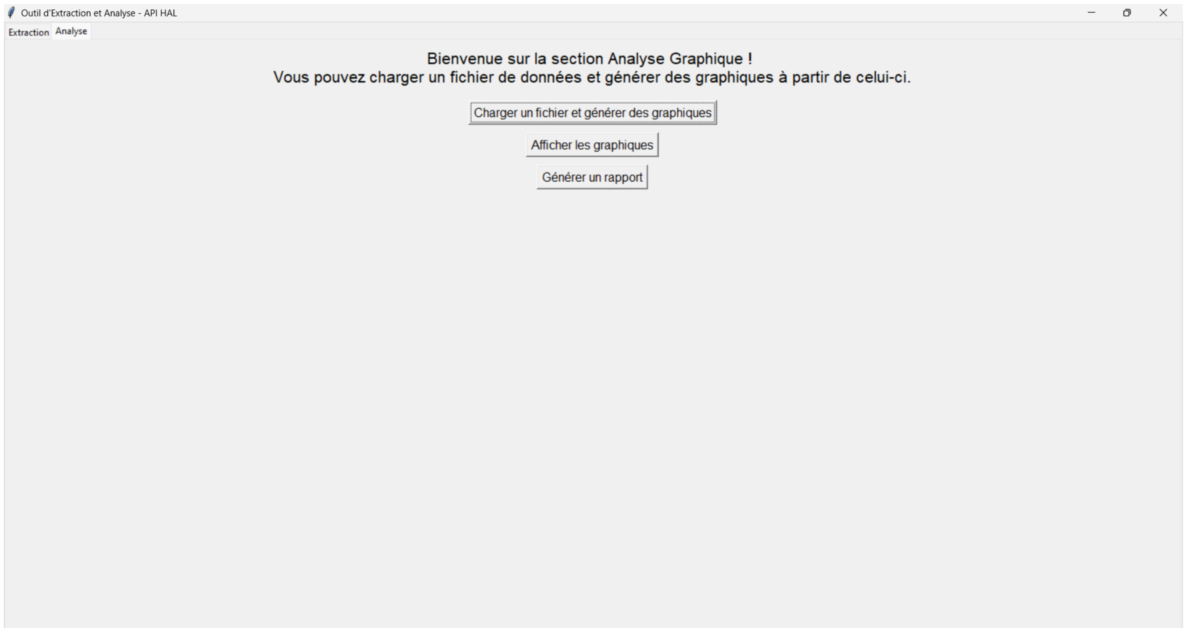
3.4.1 Automatic Visualization Generation

Simply loading an extraction CSV file triggers automatic generation of all graphics. The system simultaneously produces:

- Interactive HTML versions (`html` folder)
- Static PNG exports (`png` folder)

3.4.2 Interactive Consultation

The `Display Graphics` button automatically launches an HTML dashboard in the default browser. This dashboard, developed with `plotly`, centralizes all visualizations with advanced interactive functionalities (zoom, hover, filtering).


Outil d'Extraction et Analyse - API HAL
Extraction
Analyse

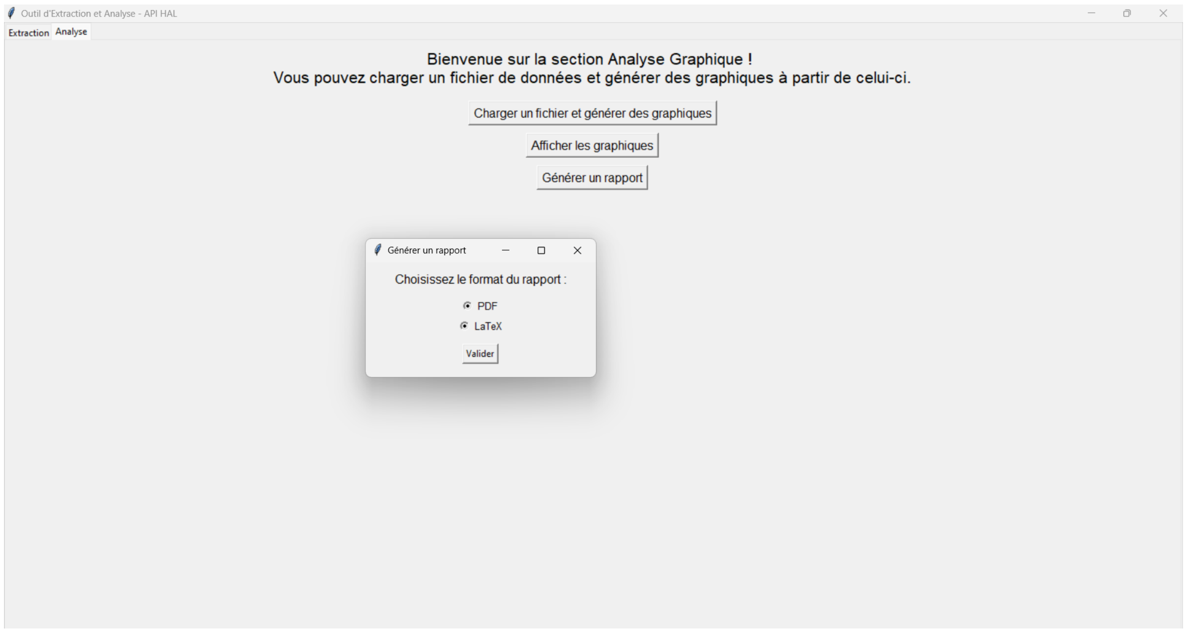
Bienvenue sur la section Analyse Graphique !
Vous pouvez charger un fichier de données et générer des graphiques à partir de celui-ci.

Charger un fichier et générer des graphiques
Afficher les graphiques
Générer un rapport

Figure 7: Consultation and Export Interface

3.4.3 Report Production

The `Generate Report` button initiates the creation of a compiled document integrating all visualizations. The user selects the desired output format:



Bienvenue sur la section Analyse Graphique !
Vous pouvez charger un fichier de données et générer des graphiques à partir de celui-ci.

Choisissez le format du rapport :

Figure 8: Report Format Selection

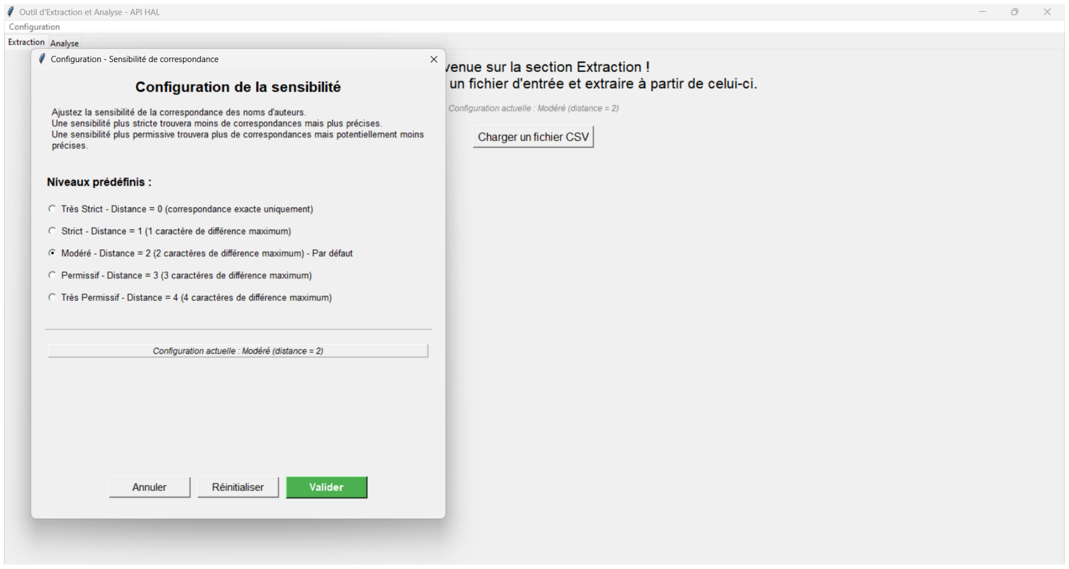
Generated reports are automatically stored in the `rapports` folder with embedded PNG images for optimal offline consultation.

3.5 Matching Sensitivity Configuration

The application integrates an advanced functionality allowing customization of the sensitivity level used during the author name and first name matching process. This configuration is useful for optimizing extraction precision based on the quality and consistency of input data.
Access to these parameters is available from the main menu bar: `Configuration > Matching Sensitivity...`.

The matching between names in the input CSV file and those in the HAL database relies on the Levenshtein distance, an algorithm that measures similarity between two character strings by calculating the minimum number of operations (insertions, deletions, or substitutions) required to make them identical. The lower this distance, the more similar the names are. This mechanism efficiently handles spelling variations, typos, accent differences, or common abbreviations.

The configuration window presents several predefined sensitivity levels, each corresponding to a specific distance value:


Outil d'Extraction et Analyse - API HAL
Configuration
Extraction
Analyse
Configuration - Sensibilité de correspondance
Charger un fichier CSV
Annuler
Réinitialiser
Valider

venue sur la section Extraction !
un fichier d'entrée et extraire à partir de celui-ci.

Configuration actuelle : Modéré (distance = 2)

Configuration de la sensibilité

Ajustez la sensibilité de la correspondance des noms d'auteurs.
Une sensibilité plus stricte trouvera moins de correspondances mais plus précises.
Une sensibilité plus permissive trouvera plus de correspondances mais potentiellement moins précises.

Niveaux prédéfinis :

- Très Strict - Distance = 0 (correspondance exacte uniquement)
- Strict - Distance = 1 (1 caractère de différence maximum)
- [✓ Modéré - Distance = 2 (2 caractères de différence maximum) - Par défaut]
- Permissif - Distance = 3 (3 caractères de différence maximum)
- Très Permissif - Distance = 4 (4 caractères de différence maximum)

Configuration actuelle : Modéré (distance = 2)

Figure 9: Sensitivity Configuration Interface

This interface offers five predefined levels:

- **Very Strict** (Distance = 0): Exact match only, no tolerance for variations
- **Strict** (Distance = 1): Tolerance for a single character difference maximum
- **Moderate** (Distance = 2): Default level, allows up to 2 character differences
- **Permissive** (Distance = 3): Extended tolerance for 3 character differences maximum
- **Very Permissive** (Distance = 4): Most tolerant level, accepts up to 4 character differences

Application example: For a name ”Müller” in the CSV file:

- *Very Strict* level: only exact ”Müller” will be found
- *Moderate* level: ”Muller”, ”Müller” will be detected
- *Permissive* level: will also include ”Miller”, ”Moller”, etc.

An information box displays the current configuration in use.
In the previous image, the ”Moderate” level is selected by default with a distance of 2. The interface also provides three action buttons: `Cancel` to close without saving, `Reset` to return to default parameters, and `Validate` to apply the new configuration before extracting data. Once validated, the new configuration will be confirmed by an informational message and will apply to all subsequent extractions.

4 Command-Line Interface Overview

The `main.py` file constitutes a complete command-line interface for extracting and analyzing scientific data from the HAL API. This approach offers a powerful alternative to the graphical interface, particularly suited for advanced users.

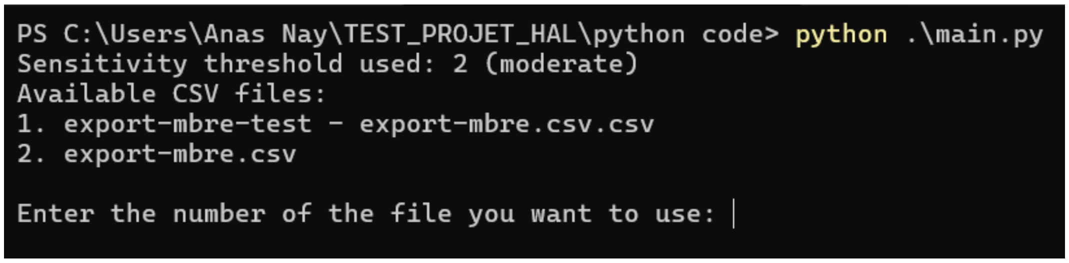
4.1 Main Functionalities

The script offers a comprehensive set of options allowing fine customization of data extraction and analysis:

- **Filtered extraction**: ability to target publications according to specific criteria
- **Sensitivity management**: configuration of author name matching threshold
- **Automatic generation**: creation of graphics and reports in a single command
- **Interactive interface**: guided selection of input files

4.2 Input File Selection

Unlike previous versions requiring manual file path modification, the current version offers an interactive interface to select the CSV file containing author names.
Upon script launch, a numbered list of available CSV files is displayed:



```
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: |
```

Figure 10: Interactive CSV File Selection

The user simply selects the number corresponding to the desired file. This improvement makes the tool more flexible and eliminates dependency on local configuration.

4.3 Filtering Options

The script accepts several optional arguments to filter extraction results:

- **–year**: specify a year range (YYYY-YYYY format)
- **–type**: filter by document type
- **–domain**: filter by scientific domain
- **–threshold**: configure name matching sensitivity (0-4)

4.4 Sensitivity Configuration

The new version integrates matching sensitivity management directly in command line via the `--threshold` argument. Available levels are:

- **0**: Very strict (exact match only)
- **1**: Strict (1 character difference maximum)
- **2**: Moderate (2 characters difference maximum) - *Default*
- **3**: Permissive (3 characters difference maximum)
- **4**: Very permissive (4 characters difference maximum)

4.5 Automatic Content Generation

Three new options enable automatic generation of visualizations and reports:

- **–graphs**: automatic graphics generation and dashboard opening
- **–reportpdf**: automatic PDF report creation
- **–reportlatex**: automatic LaTeX report generation

4.6 Utility Commands

The script also offers information commands:

- **–list-domains**: display all available scientific domains
- **–list-types**: list all supported document types
- **–list-sensitivity**: detail sensitivity levels

4.7 Usage Examples

4.7.1 Simple Extraction

```
python main.py
```

Extract all data without filters.

4.7.2 Extraction with Filters

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

Extract mathematics theses published between 2019 and 2024.

4.7.3 Extraction with Custom Sensitivity

```
python main.py —threshold 1 —domain "Computer Science"
```

Extract with strict name matching for computer science domain.

4.7.4 Extraction with Automatic Generation

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

Complete extraction with automatic graphics and reports generation.

4.7.5 Information Commands

```
python main.py --list-domains
python main.py --list-types
python main.py --list-sensitivity
python main.py -h
```

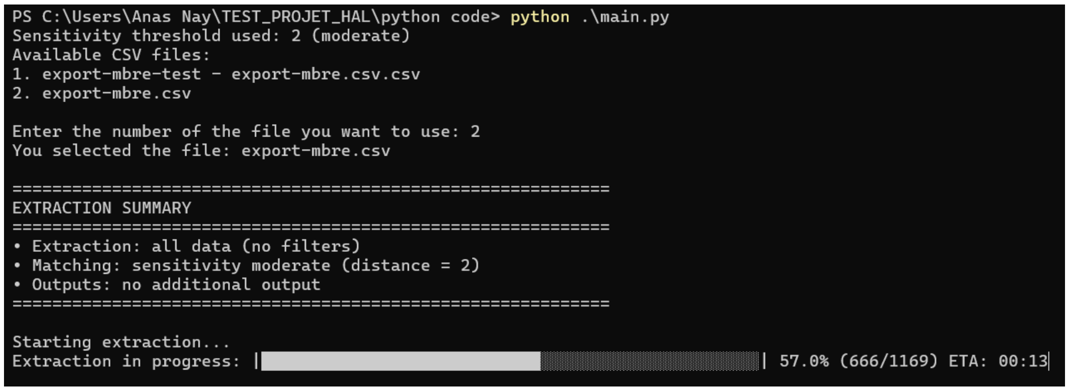
4.8 Progress Tracking and Results

The script displays a native progress bar during extraction, including:

- Completion percentage
- Number of processed elements
- Visual progress bar and Estimated time remaining (ETA)

4.9 Pre-extraction Summary

Before starting extraction, the program displays a formatted summary of selected parameters:



```
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
```

Figure 11: Pre-extraction Summary with Progress Display

4.10 Output File Organization

Results are automatically organized:

- **Extraction CSV files**: `extraction/` folder
- **HTML graphics**: `html/` folder
- **PNG images**: `png/` folder
- **Reports**: `rapports/` folder

5 Duplicate and Homonym Detection Module

This module constitutes an advanced functionality of the tool allowing automatic identification and processing of duplicate publications and homonymy cases in data extracted from HAL. This functionality relies on a method using author identifiers in the HAL database (`authIdPerson_i`) to guarantee maximum precision in detection.

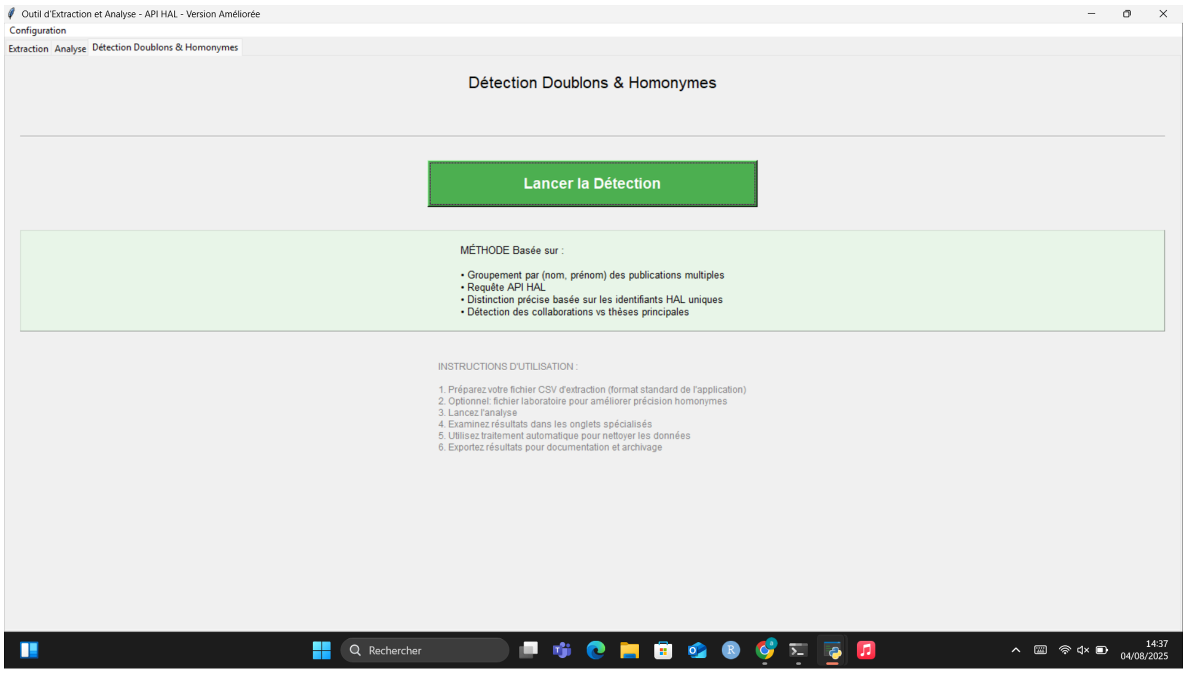
5.1 Operating Principle

The detection system is based on a multi-step algorithm:

1. **Initial grouping**: Publications are grouped by author (name, first name) pairs
2. **HAL API enrichment**: For each publication, a query is performed to the HAL API to retrieve complete metadata, notably the `authIdPerson_i` identifiers
3. **Comparative analysis**: Publications from the same author are compared according to several criteria:
   - Title similarity (default threshold: 0.8)
   - Temporal gap between publications (default threshold: 2 years)
   - Official HAL identifiers
   - Author position in the co-author list
4. **Automatic classification**: Detected cases are classified into different categories

5.2 Module Access from Graphical Interface

The detection module is accessible via the "Duplicate and Homonym Detection" tab of the main interface:


Outil d'Extraction et Analyse - API HAL - Version Améliorée
Configuration
Extraction
Analyse
Détection Doublons & Homonymes
Lancer la Détection
Rechercher
14:37
04/08/2025

Détection Doublons & Homonymes

MÉTHODE Basée sur :

- Groupement par (nom, prénom) des publications multiples
- Requête API HAL
- Distinction précise basée sur les identifiants HAL uniques
- Détection des collaborations vs thèses principales

INSTRUCTIONS D'UTILISATION :

1. Préparez votre fichier CSV d'extraction (format standard de l'application)
2. Optionnel: fichier laboratoire pour améliorer précision homonymes
3. Lancez l'analyse
4. Examinez résultats dans les onglets spécialisés
5. Utilisez traitement automatique pour nettoyer les données
6. Exportez résultats pour documentation et archivage

Figure 12: Detection Module Welcome Interface

The interface presents the main characteristics of the method used and provides a central button to launch the analysis. The process requires a CSV extraction file previously generated by the tool.

5.3 Analysis Configuration and Launch

5.3.1 Input File Selection

The analysis process begins with the selection of the CSV file to analyze. The system automatically proposes files from the `extraction` folder and asks if the user wishes to use an optional laboratory file.

5.3.2 Optional Laboratory File

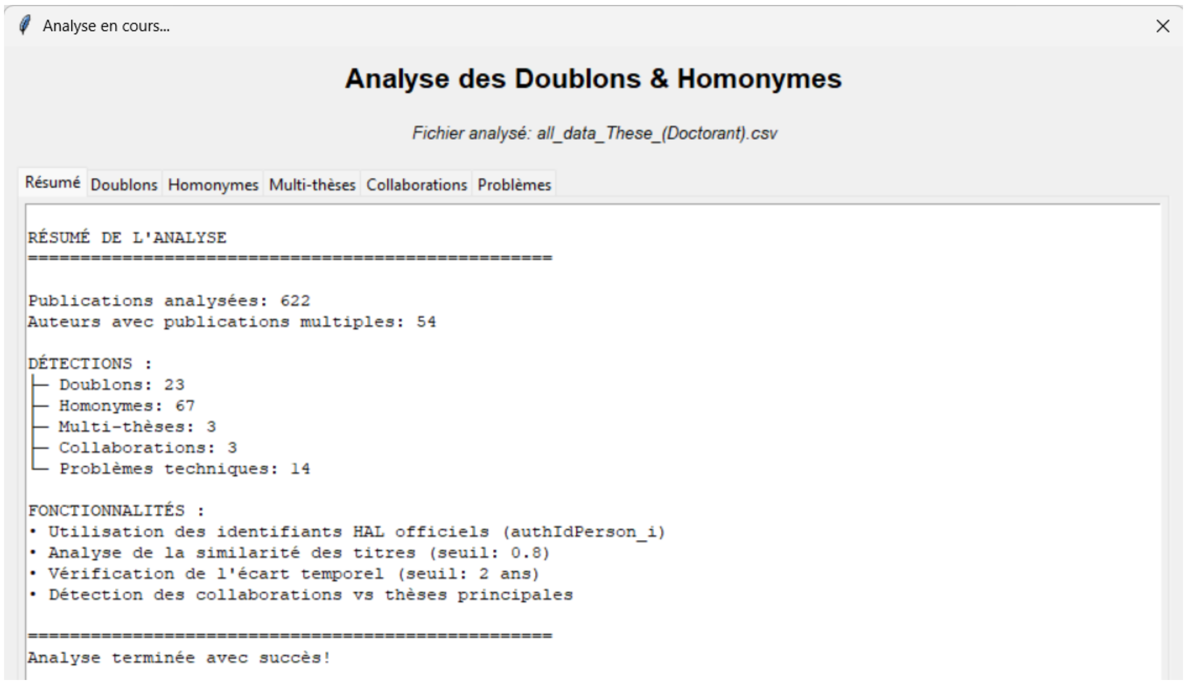
The user can optionally provide a file containing laboratory information for authors. This file must contain the columns `nom`, `prenom`, and `unite_de_recherche`. This additional information significantly improves homonym detection precision (by referring to authors' laboratory affiliations, which allows better differentiation between them).

5.4 Analysis Interface and Results

Once the files are selected, the analysis starts automatically in a dedicated interface organized into thematic tabs.

5.4.1 Summary Tab

The results are initially presented in a tab summarizing the analysis. Here is an example:


Analyse en cours...
Analyse des Doublons & Homonymes
Fichier analysé: all_data_These_(Doctorant).csv
Résumé
Doublons
Homonymes
Multi-thèses
Collaborations
Problèmes

```
RÉSUMÉ DE L'ANALYSE
==================================================

Publications analysées: 622
Auteurs avec publications multiples: 54

DÉTECTIONS :
├─ Doublons: 23
├─ Homonymes: 67
├─ Multi-thèses: 3
├─ Collaborations: 3
└─ Problèmes techniques: 14

FONCTIONNALITÉS :
• Utilisation des identifiants HAL officiels (authIdPerson_i)
• Analyse de la similarité des titres (seuil: 0.8)
• Vérification de l'écart temporel (seuil: 2 ans)
• Détection des collaborations vs thèses principales

==================================================
Analyse terminée avec succès!
```

Figure 13: Duplicate and Homonym Detection Summary

This tab presents an overview of analysis results with global statistics including:

- Total number of publications analyzed
- Number of authors with multiple publications
- Detections by category (duplicates, homonyms, multi-theses, collaborations)
- Technical information about the method used

5.4.2 Detailed Results Tabs

The different tabs present detected cases organized by category:

- **Duplicates Tab**: Publications identified as potential duplicates with similarity scores, compared titles, and publication years
- **Homonyms Tab**: Homonymy cases with differentiation criteria (different HAL identifiers, distinct scientific domains, different laboratories)
- **Collaborations Tab**: Distinction between main theses and collaborations of the same author
- **Multi-theses Tab**: Rare cases of authors with multiple theses
- **Issues Tab**: Publications without `authIdPerson_i` or incomplete metadata

5.5 Supported Detection Types

The module identifies several problem categories:

Table 1: Detection Types and Criteria

Type	**Detection Criteria**
Duplicates	Same `authIdPerson_i`, title similarity > 0.8, temporal gap < 2 years
Homonyms	Different `authIdPerson_i`, same (name, first name), distinct domains/laboratories
Multi-theses	Same `authIdPerson_i`, temporal gap > 3 years, low title similarity
Collaborations	Non-principal position in author list, different domains
Technical issues	Absence of `authIdPerson_i`, incomplete HAL metadata

5.6 Automatic Data Processing

The module offers automatic processing functionalities to clean detected data via a dedicated interface accessible after analysis.

5.6.1 Available Processing Options

Several processing levels are offered:

- **Duplicate removal**: Automatic elimination of duplicated publications (keeping the first occurrence)
- **Homonym marking**: Addition of a `Homonyme_Potentiel` column to flag ambiguous cases
- **Collaboration removal**: Elimination of collaborations while keeping only main theses
- **Multi-thesis flagging**: Marking of rare cases of multiple theses by the same author

5.6.2 Processing Results

After processing, the system generates:

- A cleaned CSV file (suffix `_nettoye.csv`)
- A detailed report of actions performed
- Comparative statistics before/after processing

5.7 Results Exportation

The module offers comprehensive export functionalities to document and archive analysis results. The user can choose the destination folder and the system automatically generates several specialized files.

5.7.1 Generated Files

The export produces several specialized files:

- `*_doublons_detecte.csv`: Detailed list of duplicates with similarity scores
- `*_homonymes_detecte.csv`: Homonymy cases with differentiation information
- `*_multi_theses.csv`: Rare cases of multiple theses per author
- `*_collaborations.csv`: Main theses/collaborations distinction
- `*_resume_detecte.txt`: Complete analysis report with methodology

5.8 Command-Line Usage

The detection module is also accessible via the command-line interface with the `--analyse` argument:

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

This approach offers the same functionalities in an interactive textual environment:

1. Interactive selection of the CSV file to analyze
2. Optional configuration of the laboratory file
3. Analysis with real-time results display
4. Integrated processing and export options

5.8.1 Command-Line Session Example

```
# Launch analysis
python main.py --analyse

# File selection (interactive interface)
Available files in 'extraction/':
1. all_data_These_Doctorant.csv
2. publications_2020-2024.csv
Select a file (1-2): 1

# Laboratory file configuration
Use a laboratory file? (y/n): y
Laboratory file selected: laboratoires.csv

# Analysis in progress with progress bar
Analysis in progress... (patience required - HAL API querying)
Analyzing Dupont Jean (3 publications) - 15/120...

# Detailed results display
============================================================
ANALYSIS RESULTS
============================================================
Publications analyzed: 450
Authors with multiple publications: 120
Duplicates detected: 23
Homonyms detected: 8
Collaborations detected: 15
```

5.9 Limitations and Considerations

5.9.1 HAL API Dependency

Detection precision depends on HAL metadata quality:

- Publications without `authIdPerson_i`: processing by title similarity only
- Incomplete metadata: classification in "technical issues"
- HAL data evolution: possible variations over time

5.9.2 Special Cases

Some cases require manual validation:

- Author name changes (marriage, etc.)
- International collaborations with spelling variations
- Publications with metadata errors on the HAL side

5.10 Conclusion

The duplicate and homonym detection module constitutes a powerful tool for improving bibliometric analysis data quality. By combining the use of official HAL identifiers with advanced heuristics, it enables reliable automatic detection of most problematic cases while providing appropriate processing solutions.
The integration of this module into the global tool ensures a complete processing chain, from initial extraction to final analysis, through a data cleaning and validation phase that significantly improves the relevance of obtained results, and consequently the statistics generated thereafter.

6 Appendix

6.1 Expected CSV File for Extraction

To download and view the type of CSV file required for extraction, please use the following link:

Download the CSV file here

6.2 Description of the Obtained CSV File

The CSV file obtained after data extraction, whether via the application interface or by executing the `main.py` file, presents a uniform structure. The file will contain the same information regardless of the method used for extraction. Each line in the CSV file represents a publication. Here is a description of the typical columns found in this file:

Table 2: CSV File Column Description

Column	**Description**
Nom	The author's last name.
Prénom	The author's first name.
IdHAL de l'Auteur	HAL identifier of the author.
IdHAL des auteurs de la publication	HAL identifiers of the publication authors.
Titre	The publication title.
Docid	The publication identifier.
Année de Publication	The publication year.
Type de Document	The document type, for example article, thesis, etc.
Domaine	The scientific domain of the publication.
Mots-clés	The keywords associated with the publication.
Laboratoire de Recherche	The laboratory or research center associated with the publication author.