Ptéryx

Index

1. Introduction

The Evolution of Scientific Research and the Need for Automation

Until recently, scientific research relied on physical documents—researchers had to flip through pages, manually note relevant information, and handle large volumes of paper. This process was time-consuming, labor-intensive, and prone to human error.

The introduction of the PDF format in 1993 revolutionized how we store and access information. Documents began to be digitally archived, simplifying organization and eliminating stacks of paper. However, even with this digital shift, manually searching through hundreds or thousands of PDFs remains extremely tedious.

Today, most academic articles and books are available online, drastically reducing the circulation of physical materials. In Brazil, the impact is stark: over the last decade, 17% of newsstands and 29% of bookstores have closed, and between 2015 and 2020, at least 764 public libraries were lost (National Public Library System). This forces researchers to depend on electronic devices for reading—extended screen time can cause visual fatigue and other eye problems.

The Challenges of Manual PDF Search

Manually searching digital documents involves opening files, reading each page, taking notes, and repeating the process continuously. This method, besides being extremely time-consuming, increases the probability of errors, as prolonged reading fatigue can lead to the omission of relevant data. Studies indicate that reading on backlit screens requires greater visual effort, impacting the accuracy of analyses and making the process less efficient.

Beyond cognitive strain, the continuous repetition of these tasks physically and mentally exhausts the researcher. Considering that reading a 5-page PDF takes an average of 5 minutes, a professional would need 8 hours to review approximately 100 documents. To screen 3,000 files, approximately 30 days of uninterrupted work would be required. Why perform this task manually when software can execute it in a few minutes, with greater accuracy and without interruptions?

Limitations of Windows Search and the Pteryx Solution

Microsoft Windows allows searching for information within multiple files, but this functionality has limitations: it only permits searches for one term at a time, does not generate detailed reports, and does not automatically organize files. This makes manual screening slow and ineffective for scientific research.

Pteryx was developed to solve this problem. Created by a zoology researcher, it offers a robust and automated solution for analyzing scientific documents in PDF format. The software preprocesses files, performs automated searches, and generates detailed reports, indicating where each piece of information was found and in what context. This makes document screening faster, more accurate, and efficient, allowing researchers and academics to focus on data analysis instead of wasting time on repetitive tasks.

Automating this stage is essential to keep up with the increasing digitization of science and ensure that research continues to advance without the obstacles of manual work.

---

2. Installation and System Requirements

2.1 Minimum Requirements

- **Operating System**: Windows 10 or higher
- **Processor**: Intel Core i5 or equivalent
- **RAM**: 8GB or higher
- **Disk Space**: 2GB free for installation
- **Minimum resolution**: 1280 x 720
- **Required Software**: Microsoft Excel, or equivalent, installed (for report manipulation)
- **Required Libraries**: The software includes all dependencies in the executable, with no need for manual installation.

2.2 Installation

1. Download the Pteryx installer from Pteryx website.
2. Run the **Pteryx_installer.exe**, choose the installation drive, and enter your e-mail. The installer will download the application and create the necessary folder structure.

2.3 languages

Pteryx automatically detects your system language and will run in one of the following: English, Portuguese, Spanish, Galician, Bulgarian, French, Russian, German, Japanese, or Mandarin. For any other language, the interface defaults to English.

---

3. User Interface

3.1 Pre-registration

Before purchasing and registering the program, it will work in demo mode, showing a “register now” button on the main screen (Fig. 1). In demo mode, pre-processing does not work, and searches are limited to three in the “Search by species” and “word or phrase” modes.

After completing the three searches, the program will go directly to the registration screen (Fig. 2).

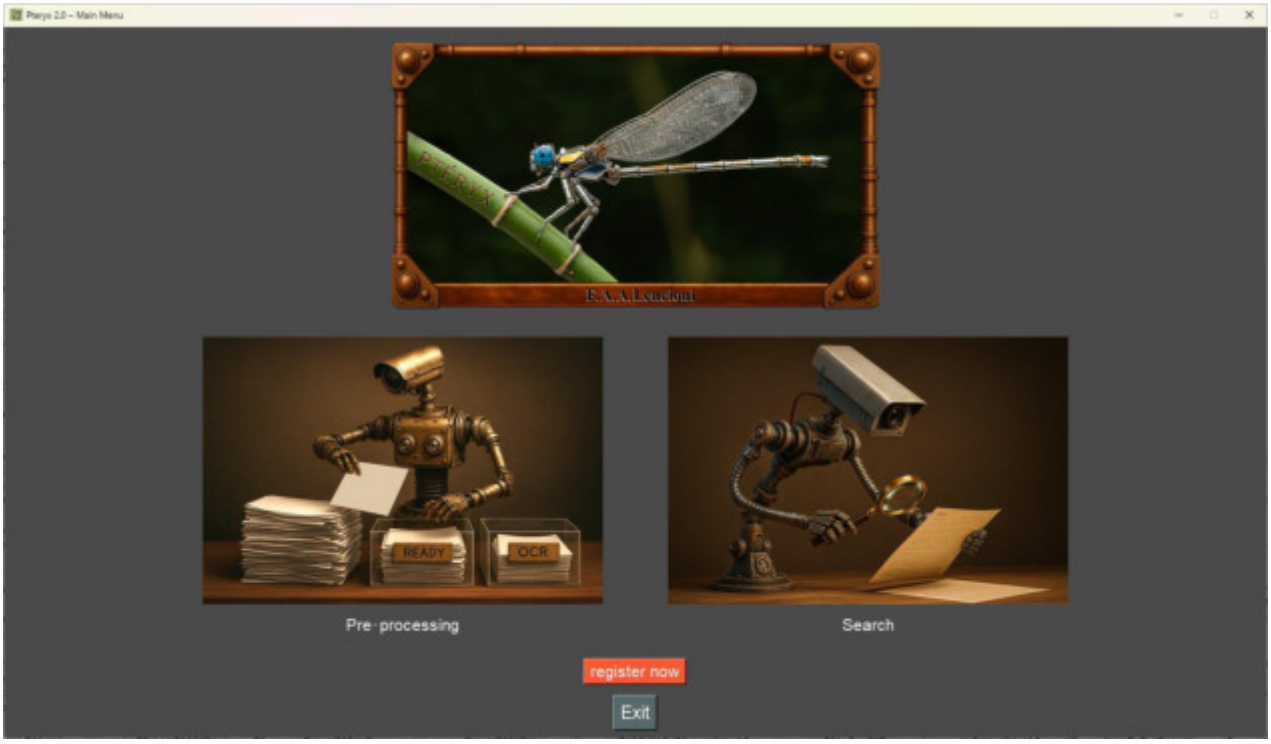

Fig. 1 - Main Menu (pre-registration screen).

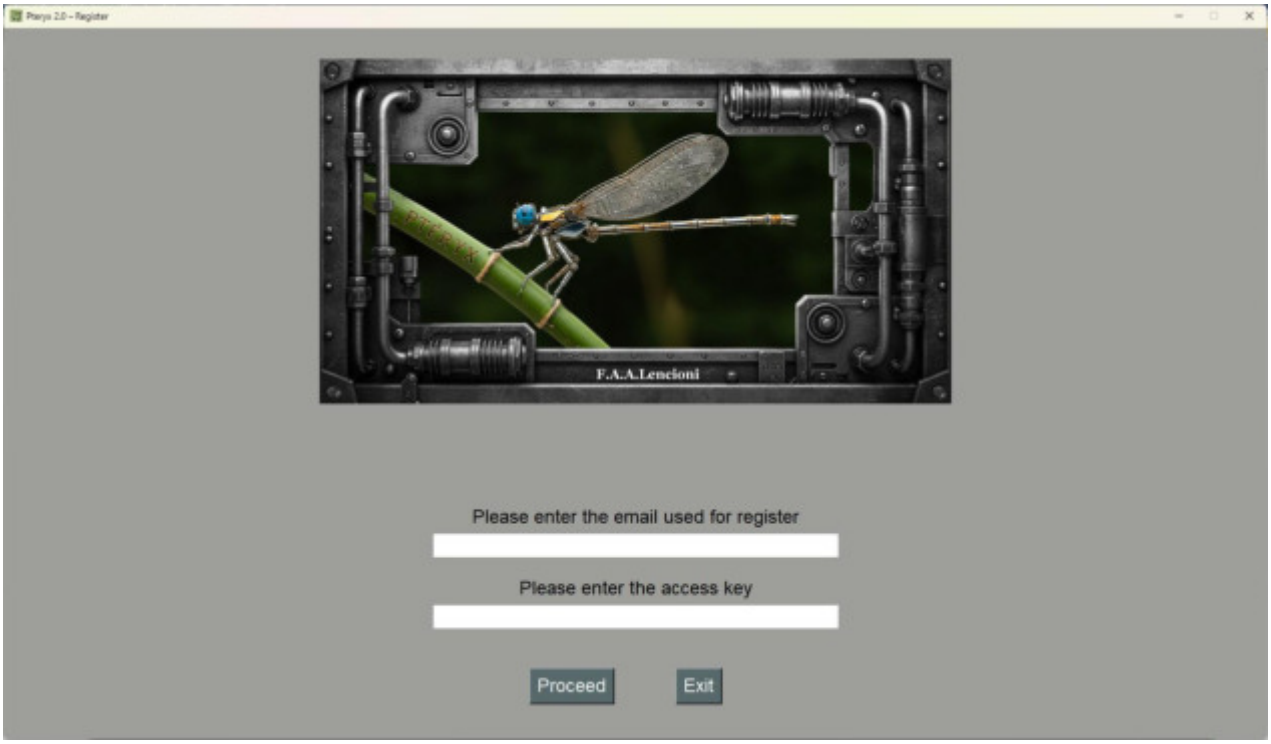

Fig. 2 - Register screen.

3.2 License acquisition

To purchase the license, visit the website **www.pteryx.com.br** and as soon as payment is made, you will receive in the registered email the access key to unlock the software.

3.3 Post registration - Home Screen

After registration, you arrive at the main screen (Fig. 3). From here you can select one of two modules:

- **Pre-processing** (PDF Pre-processing)
- **Search** (Search Module)

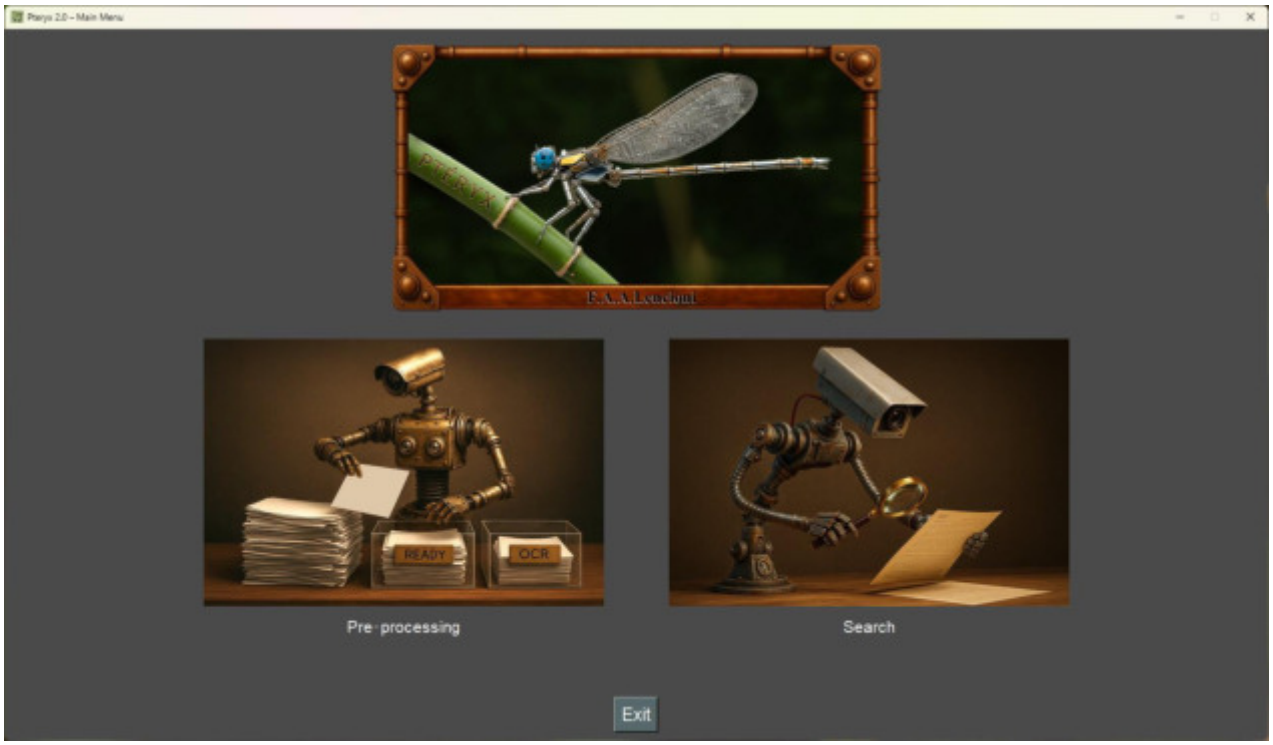

Fig. 3 – Main Menu.

4. PDF Pre-processing Module

The software is designed to search PDFs and generate detailed reports of these searches. To do so, the PDFs must be readable and searchable, error-free, and organized in a predetermined location.

The pre-processing system does this in two steps:

Step 1: Searches for and copies all PDFs from a folder or disk to the location where they will be processed (the user has the option of manually placing the PDFs in the **1_PDFs_to_Process** folder).

Step 2: Screens all PDFs, looking for errors, duplicates, and image-based PDFs (which are moved to the **2_PDFs_to_OCR** folder).

4.1 Organizing Files

- Click the **"Pre-processing"** button on the main screen.
- You can manually copy and paste PDFs into the **1_PDFs_to_Process** folder or click **"Gather PDFs"** and specify which folders or drives the program should collect PDFs from.
- The software will only copy your files; your original structure will be maintained intact.
- If you choose multiple folders that contain copies of the same PDF, they will be saved with the suffix (1). If more than one copy exists, they will be saved with sequential numbers (2, 3, 4, etc.).

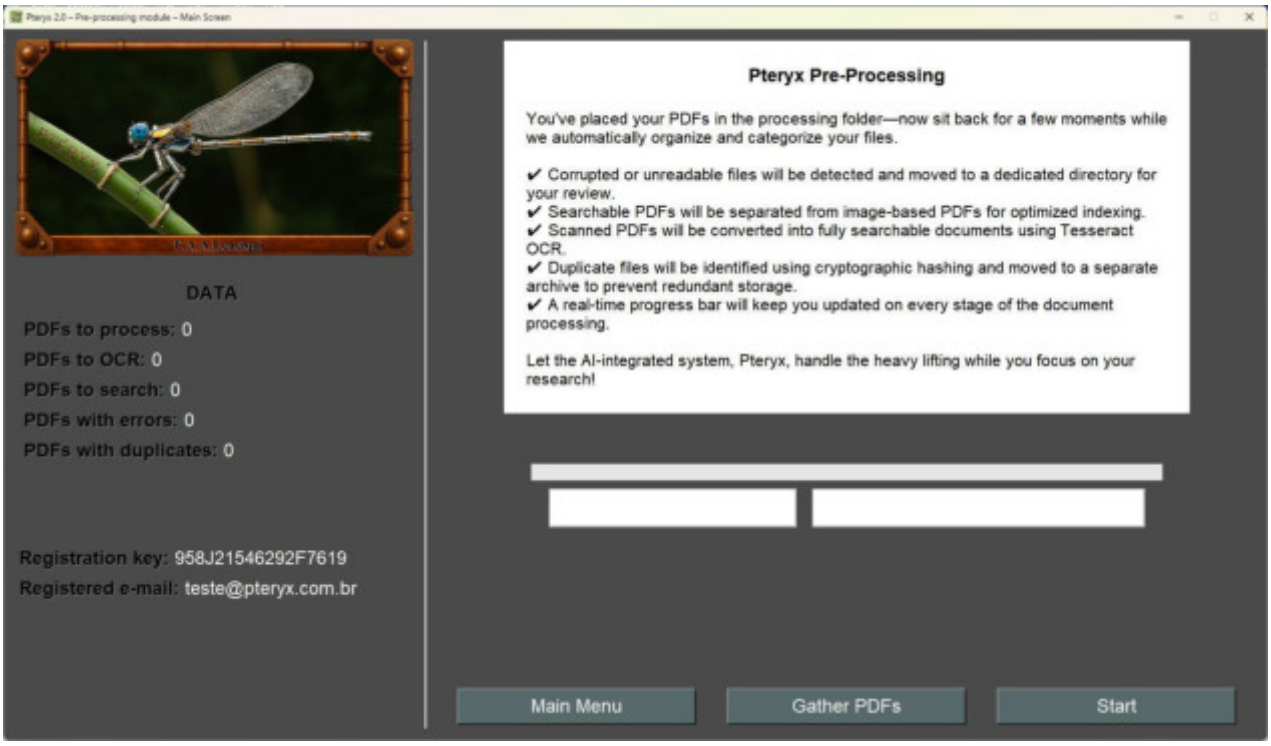

Fig. 4 – PDF Pre-processing.

- Click the **"Gather PDFs"** button.

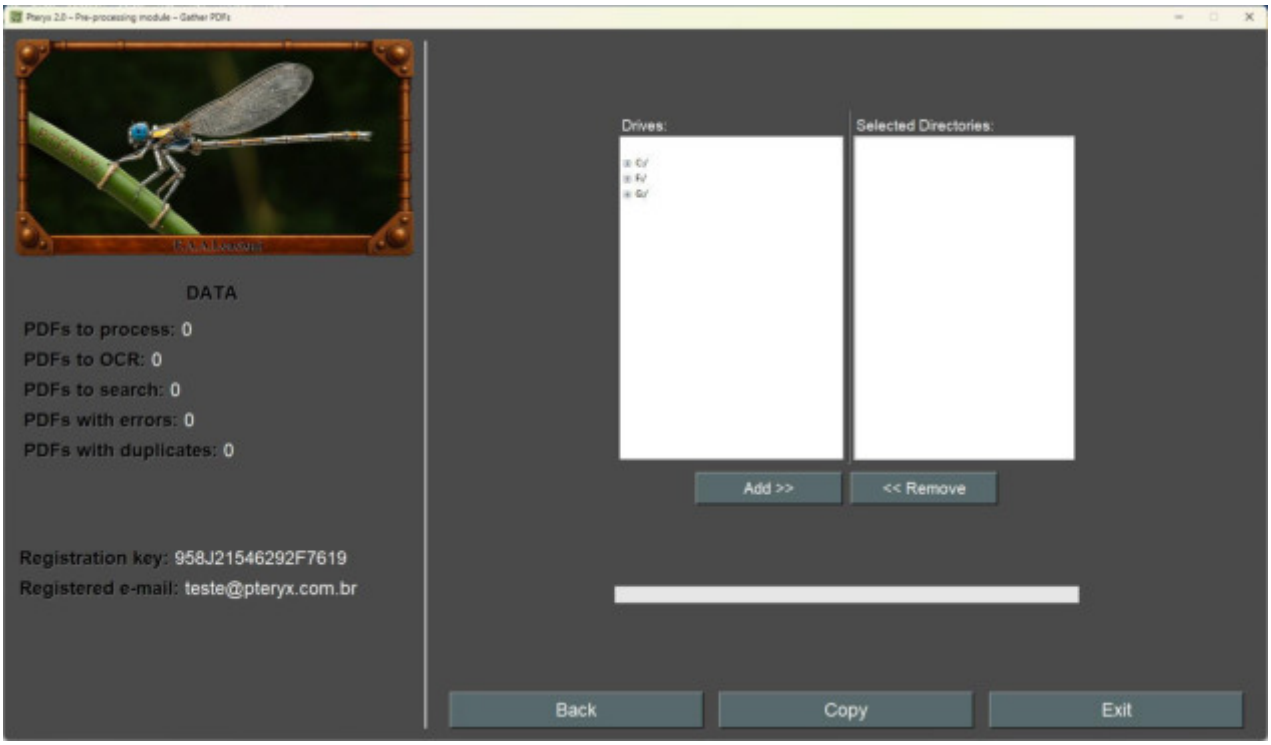

Fig. 5 – Gather PDFs, choosing folders.

- Choose the folders or drives from where the PDFs will be gathered and add them to the **"Selected directories"**.
- Press **Copy** and all PDF files from the selected folders or drives will be copied to **1_PDFs_to_Process**.

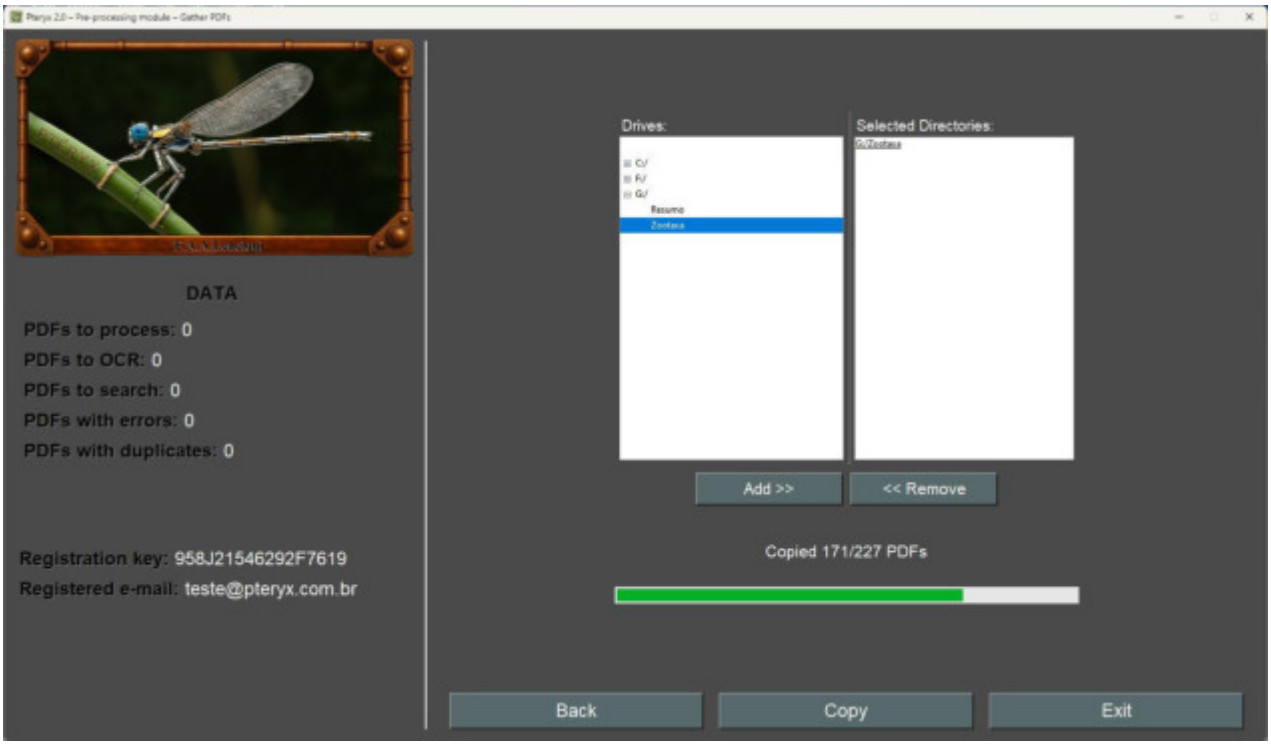

Fig. 6 – Gather PDFs, copying files.

4.2 Processing

Once all PDFs are in the **1_PDFs_to_Process** folder, go back to the pre-processing main menu and click **"Start"**. The system will process all files, checking for:

- Corrupted or error files → moved to **4_PDFs_with_Errors** (Fig. 7).
- Duplicate file copies → moved to **5_PDFs_with_duplicates** (Fig. 8).

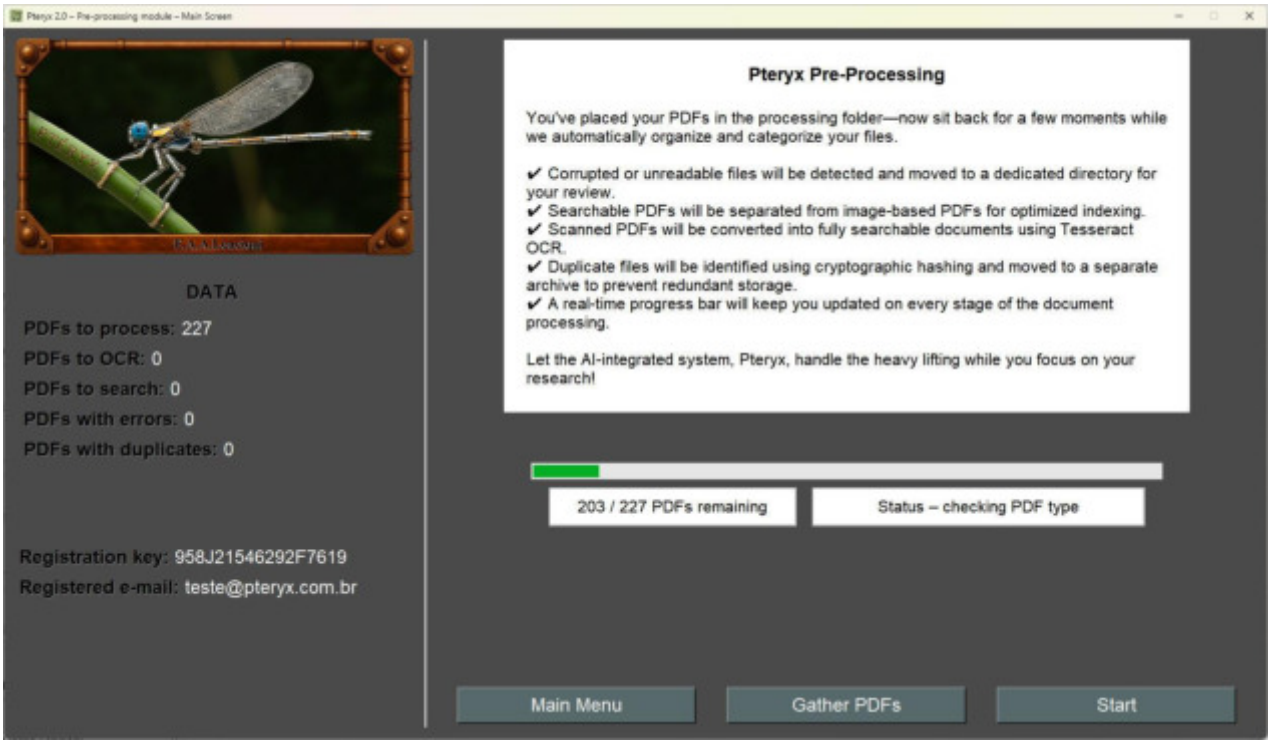

Fig. 7 – Checking file type.

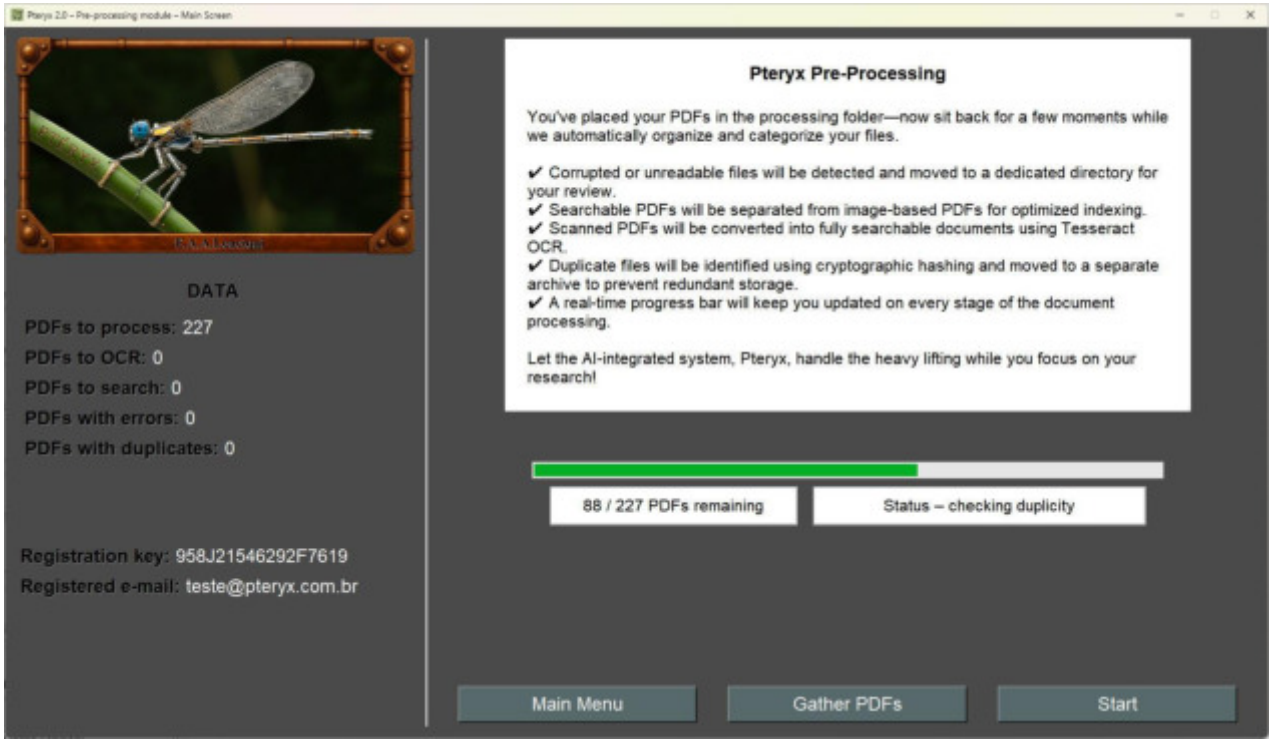

Fig. 8 – Checking for duplicate files.

- Image-based PDFs requiring OCR → moved to **2_PDFs_to_OCR.**
- Ready-to-search PDFs → moved to **3_PDFs_to_search.**

4.3 OCR Execution

Once the pre-processing is completed, if there are files that require OCR (image-type PDFs), the system will display the number of files and ask whether you want to execute OCR (Fig. 9). If you choose **Yes**, simply wait for all the files to be processed (Fig. 10).

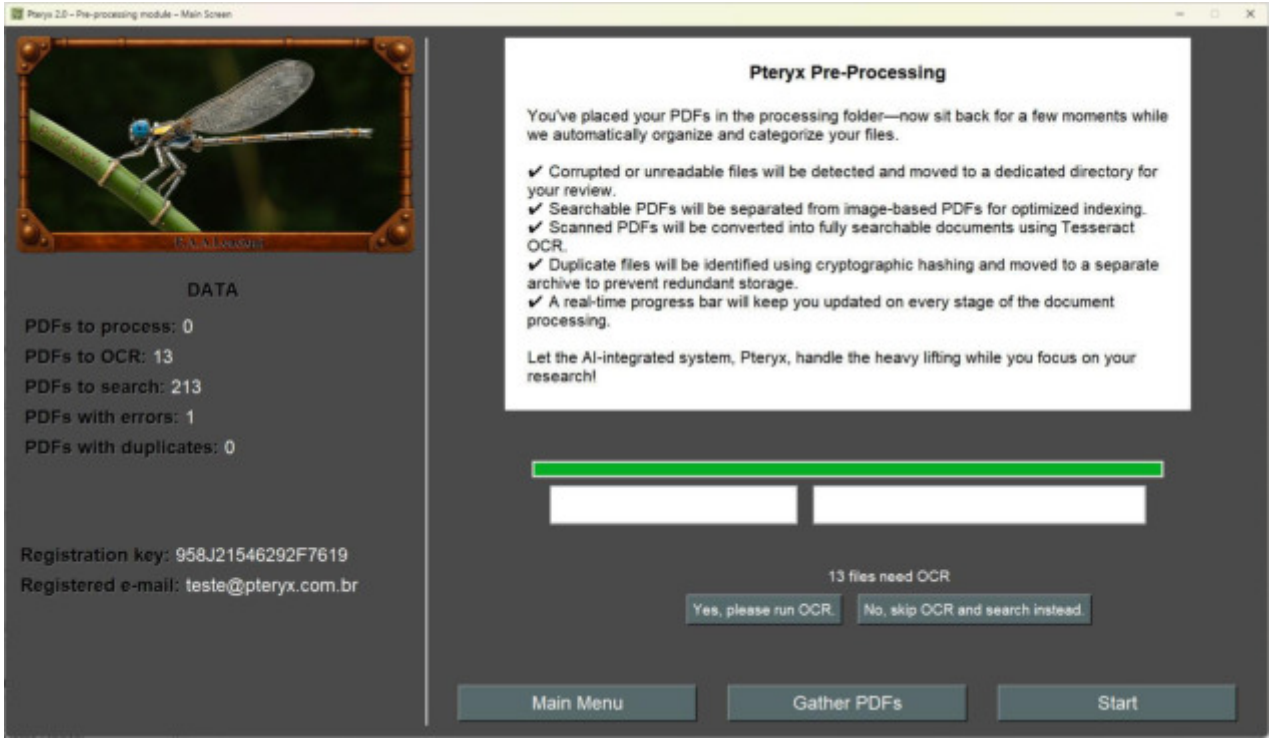

Fig. 9 – OCR prompt.

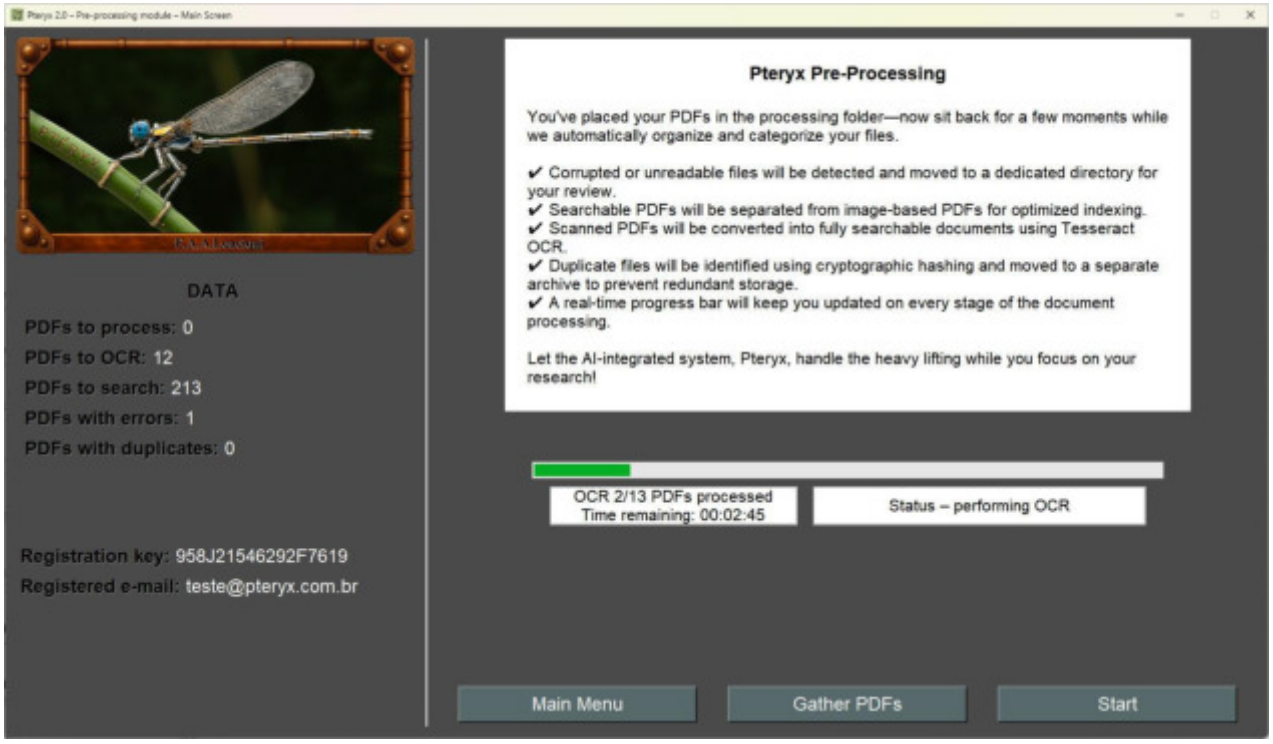

Fig. 10 – Running OCR.

4.4 Finalizing

After the OCR process is completed, the files will be moved to the **3_PDFs_to_search** folder, and a report of all processed PDFs will be displayed (Fig. 11).

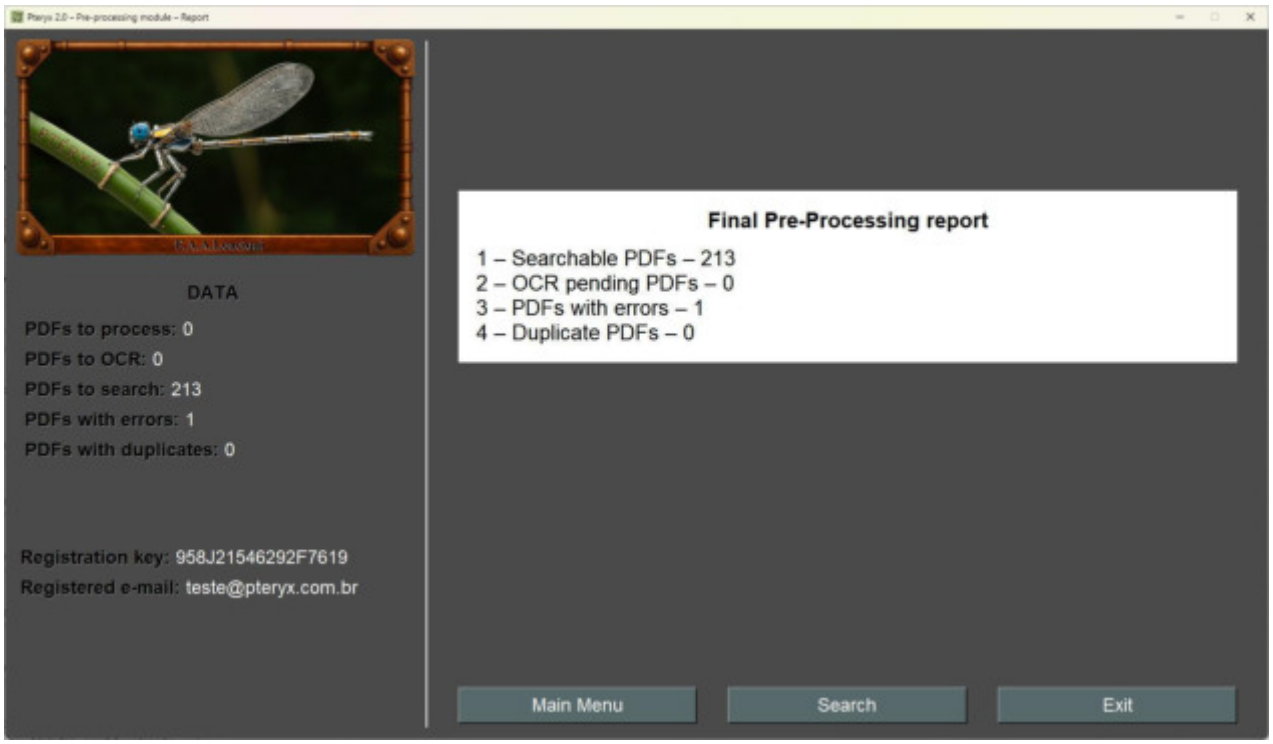

Fig. 11 – Final Pre-processing Report.

4.5 Proceed

Now, return to the Main Menu to perform your searches.

5. Search Module

This is the main module, offering three search methods. Each generates detailed, easy-to-review reports (Fig. 12).

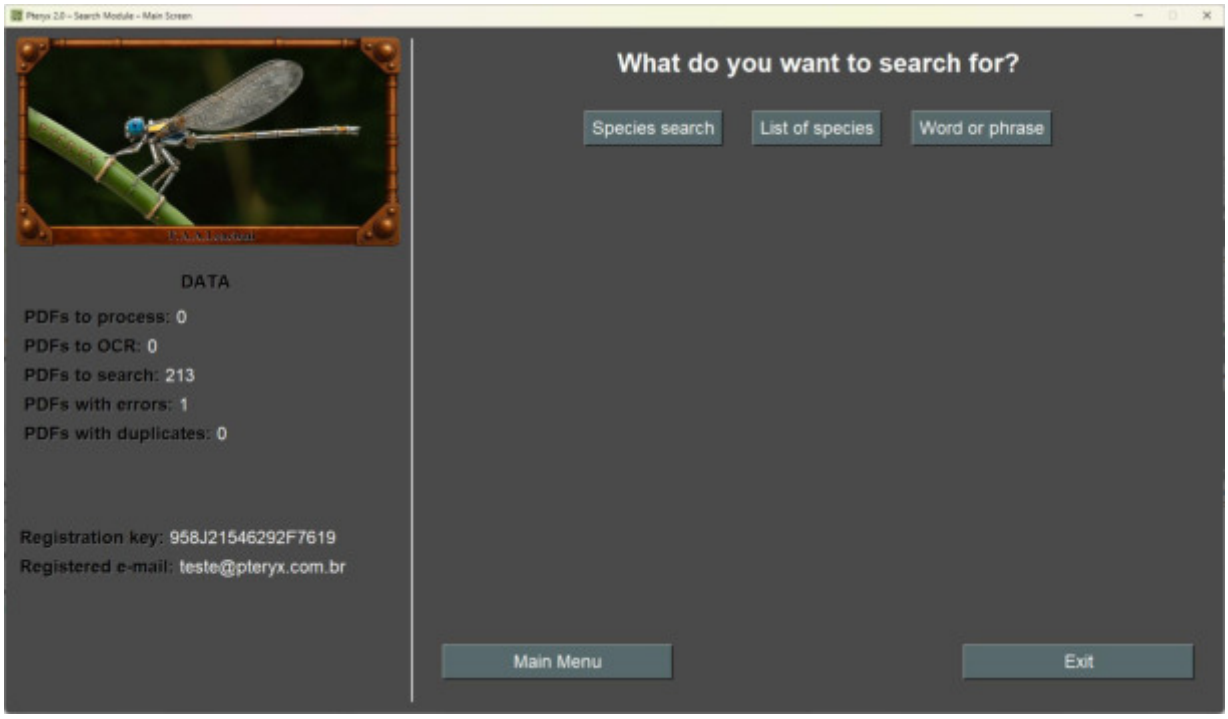

Fig. 12 – Search module.

5.1 Search by Species

This function searches all PDFs stored in the **3_PDFs_to_search** folder for occurrences of a specific species. The search can be performed with the genus input (Fig. 13) or without it (Fig. 14).

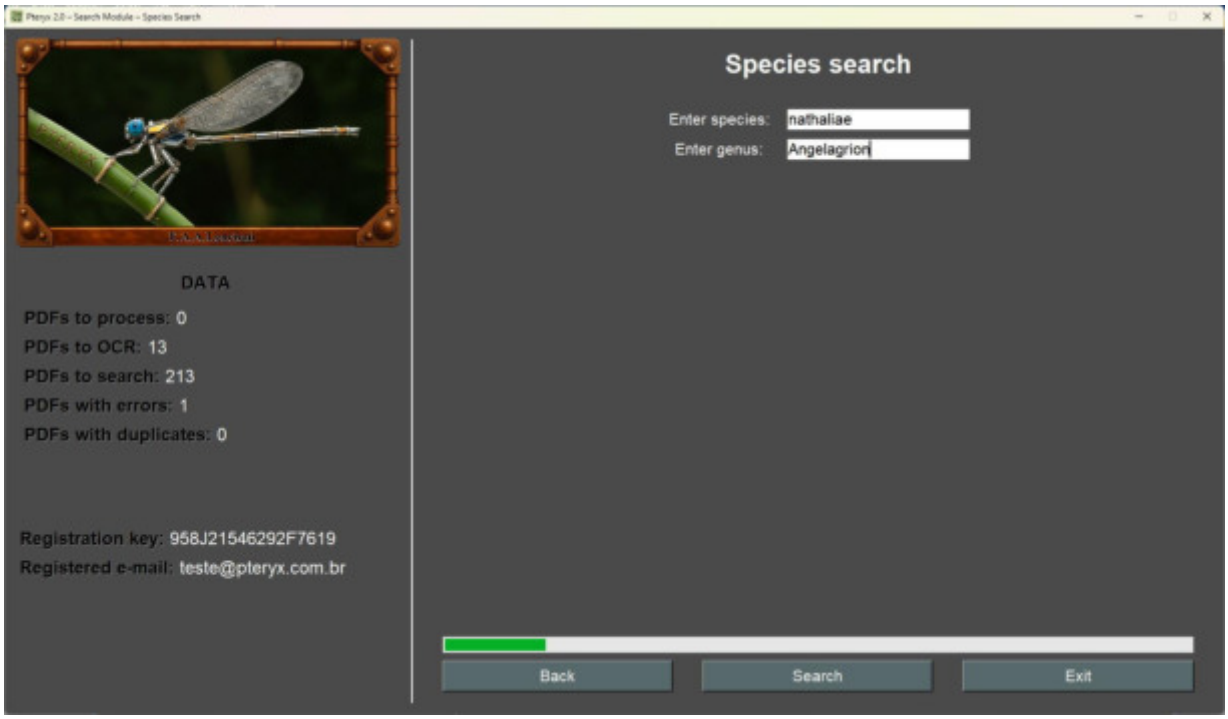

Fig. 13 – Search by species, with genus.

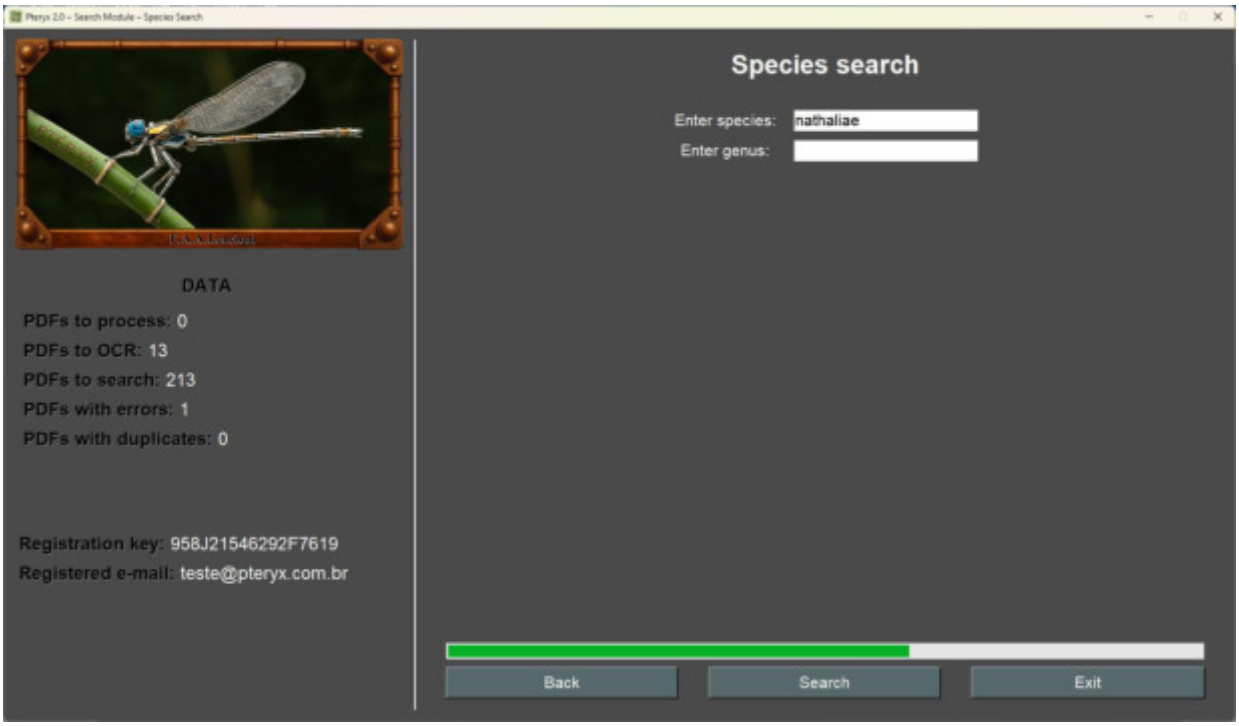

Fig. 14 – Search by species, without genus.

5.2 Search by Species List:

This search is based on an **Excel file** (Data.xlsx – Fig. 15) containing multiple species.

In this search mode, the user can search for 2 to hundreds of species at once, which greatly facilitates working with large volumes of data, such as when writing systematic reviews and theses. We usually gather hundreds and hundreds of documents, and the analysis is always slow, subject to errors, tiring and time-consuming.

With this mode, you create a simple spreadsheet and let the system work for you. If you are searching in hundreds of documents, you can let the system work while you sleep or do other activities and reap the rewards when you wake up or return from work.

You need to create the spreadsheet following the following criteria:

- The file must contain a **sheet** (named **"Search"**, please pay attention that the first letter is capitalized) with the following five columns: **Species, Genera, Key, Full name** and **Name with abbreviation.**
- There is a sample file in the package (**Data.xlsx**) already formatted correctly; you only need to fill in the data and reference it when requested.
- The **"Full name"** and **"Name with abbreviation"** fields can be filled in manually, but if left empty, the system will automatically complete them.
- The **key** can be any word, but it must not be a subset of another key (unless connected by a hyphen). Example: If **Coenagrionidae** and **Non Coenagrionidae** are entered, the system will interpret the second as part of the first. However, if they are written with a hyphen (**Non-Coenagrionidae**), the system will treat them as distinct words.
- Some names are routinely used to name species, some unfortunately belong to distinct genera but have identical initials, such as: *Pantala flavescens* (a dragonfly), *Perca flavescens* (a fish), *Papiliotrema flavescens* (a fungus), *Parachalastinus flavescens* (a beetle), *Pulvinaria flavescens* (Homoptera), etc, all of them will be abbreviated as "***P. flavescens***" and therefore if they are found in the searched PDFs they will return as positives, to facilitate the screening of the data obtained, there is a "context" column in the generated spreadsheet that facilitates the finding of these false homonyms.

	A	B	C	D	E	F
1	Species	Genera	Key	Full name	Name with abbreviation	
2	Species_1	Genera_1	interest	Genera_1 Species_1	G. Species_1	
3	Species_2	Genera_2	interest	Genera_2 Species_2	G. Species_2	
4	Species_3	Genera_3	interest	Genera_3 Species_3	G. Species_3	
5	Species_4	Genera_4	interest	Genera_4 Species_4	G. Species_4	
6	Species_5	Genera_5	interest	Genera_5 Species_5	G. Species_5	
7	Species_6	Genera_6	interest	Genera_6 Species_6	G. Species_6	
8	Species_7	Genera_7	interest	Genera_7 Species_7	G. Species_7	
9	Species_8	Genera_8	interest	Genera_8 Species_8	G. Species_8	
10	Species_9	Genera_9	interest	Genera_9 Species_9	G. Species_9	
11	Species_10	Genera_10	interest	Genera_10 Species_10	G. Species_10	
12	Species_11	Genera_11	Proto	Genera_11 Species_11	G. Species_11	
13	Species_12	Genera_12	Proto	Genera_12 Species_12	G. Species_12	
14	Species_13	Genera_13	Proto	Genera_13 Species_13	G. Species_13	
15	Species_14	Genera_14	Proto	Genera_14 Species_14	G. Species_14	
16	Species_15	Genera_15	Proto	Genera_15 Species_15	G. Species_15	
17	Species_16	Genera_16	Proto	Genera_16 Species_16	G. Species_16	
18	Species_17	Genera_17	Proto	Genera_17 Species_17	G. Species_17	
19	Species_18	Genera_18	Proto	Genera_18 Species_18	G. Species_18	
20	Species_19	Genera_19	Proto	Genera_19 Species_19	G. Species_19	
21	Species_20	Genera_20	Proto	Genera_20 Species_20	G. Species_20	
22						
23						

Search

Fig. 15 –Species list, example of file.

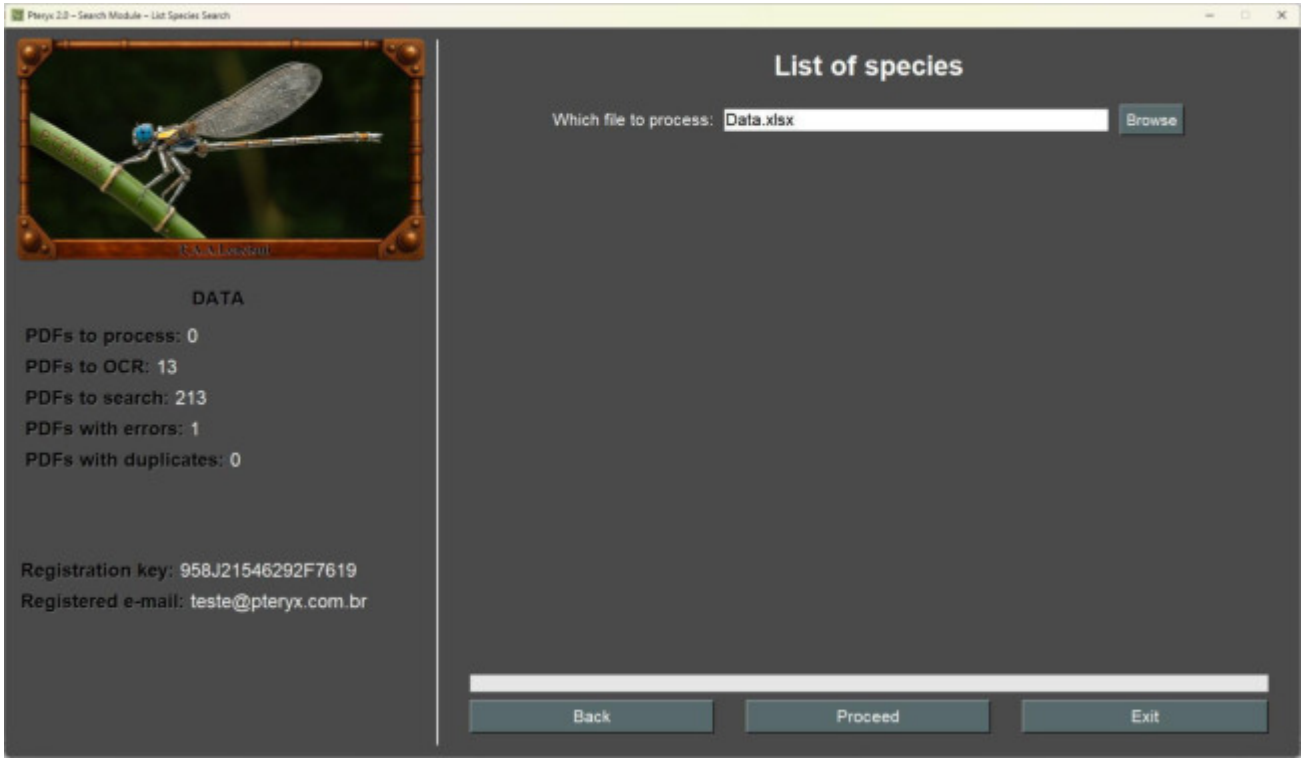

Fig. 16 – Search by species list, select the file.

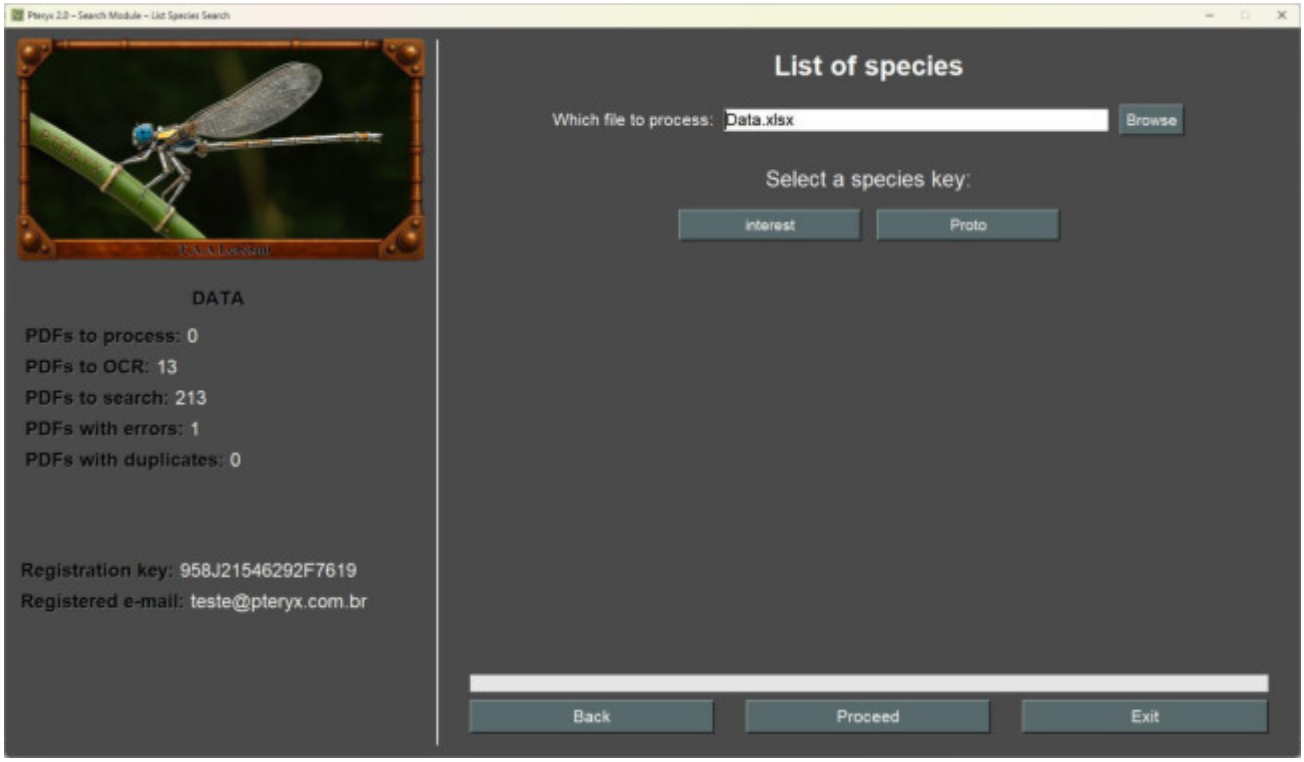

Fig. 17 – Search by species list, select the key to be searched; this key groups the species list in the **.xlsx** file.

5.3 Search by Terms or Phrases

This function locates specific **keywords or phrases** within the documents. The search is conducted across all PDFs stored in the **3_PDFs_to_search** folder. The term or phrase can be **up to 150 characters** (Fig. 18).

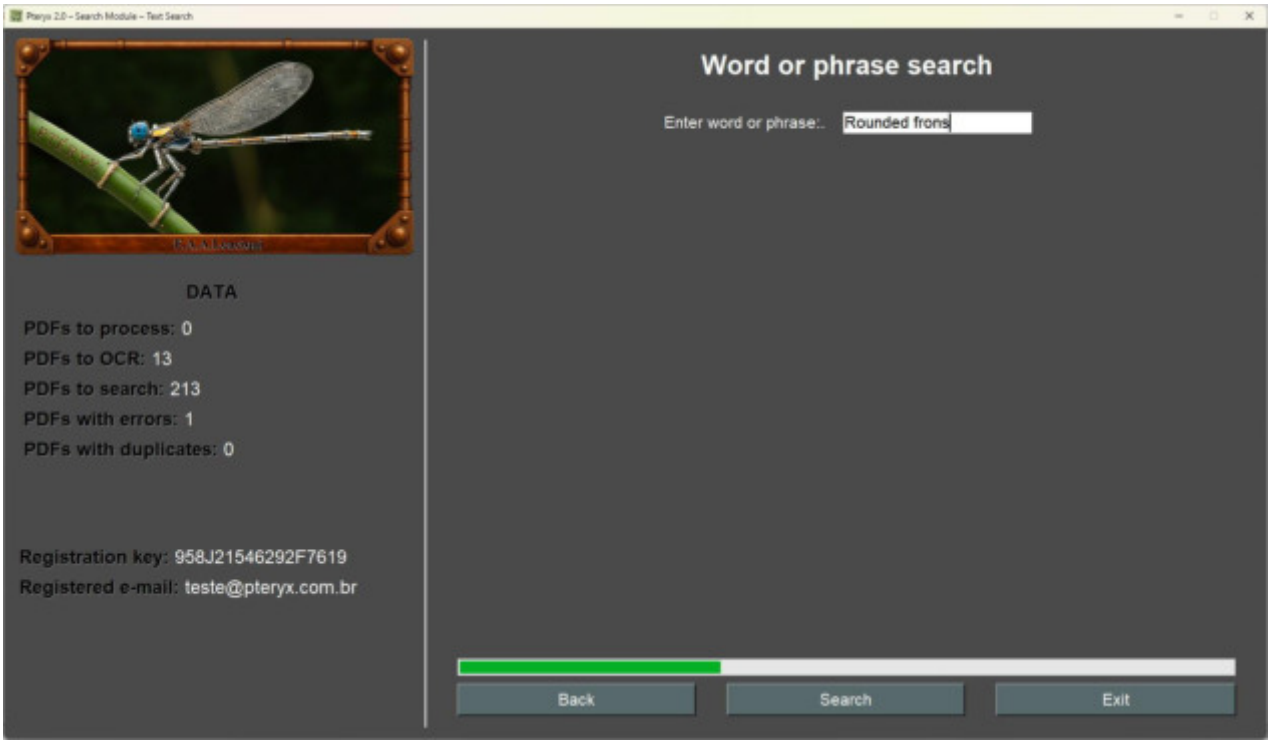

Fig. 18 – Search by word or phrase (limited to 150 characters).

6. Export and Reports

After a search, the results are **exported**:

- **Excel spreadsheets (.xlsx)** – located in the **6_PDFs_with_target_data** folder (Fig. 19).
- **Files organized by species or keyword** – stored in directories named after the searched species within **6_PDFs_with_target_data**.

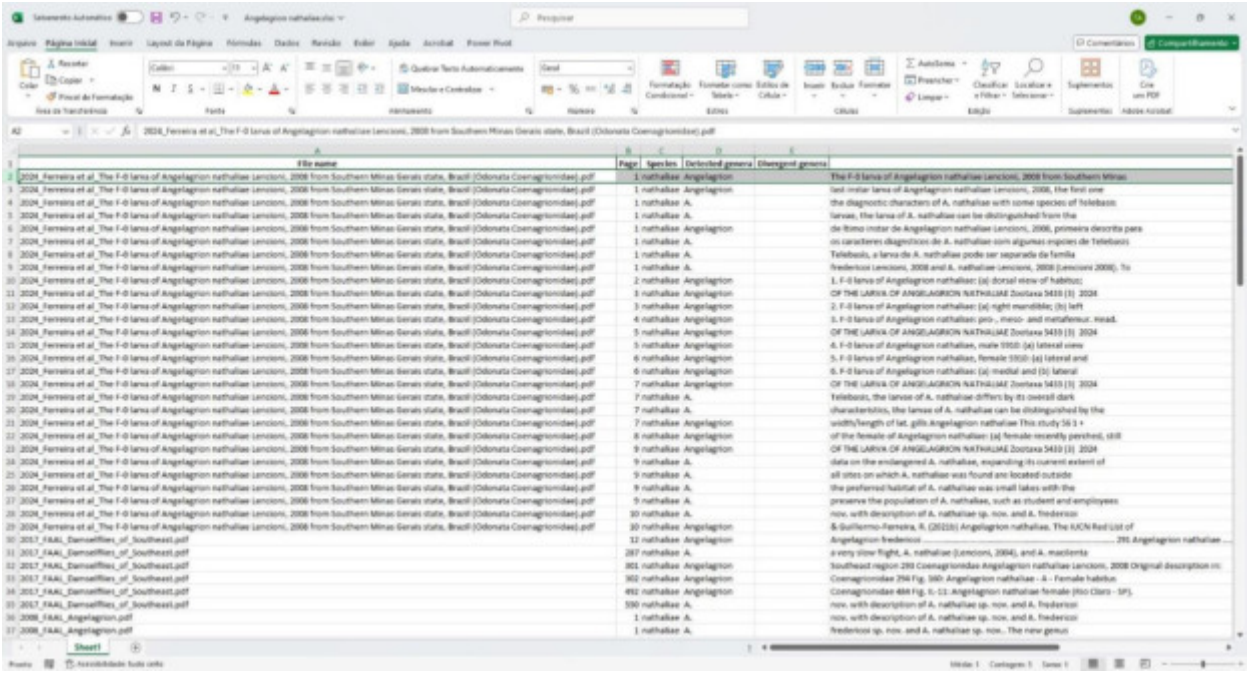

Fig. 19 – Example of a report of the located data.

7. Troubleshooting

Error: The software does not start
✓ Ensure all software files are in the same folder.
✓ Try running the program as an Administrator.

Error: No search results found
✓ Check that the PDFs were correctly processed in the Pre-processing module.
✓ Verify that the search terms are spelled correctly.

---

8. Contact and Support

For questions and technical support, contact via email: **sac@pteryx.com.br**

---

9. Glossary

OCR (Optical Character Recognition)
Technology that converts images containing text (such as scanned PDFs) into searchable and editable text.

PDF (Portable Document Format)
A file format used to present documents independently of software, hardware, or operating system.

Pre-processing
Initial stage where PDFs are organized, checked, duplicates removed, and prepared for searches or OCR.

Hash (Hashing)
Method to identify unique files by generating a numerical sequence based on the file's content, used to detect duplicates.

Duplicates
Identical files that are automatically identified and separated to avoid redundant processing.

Folder “1_PDFs_to_Process”
Directory where the user should place the original PDFs to start processing in Pteryx.

Folder “2_PDFs_to_OCR”
Folder where image-based PDFs that require OCR are stored.

Folder “3_PDFs_to_search”
Folder where processed PDFs ready for search are stored.

Search by Species
Search for occurrences of a specific species within the processed PDFs.

Search by Species List
Simultaneous search for multiple species based on a formatted Excel spreadsheet.

Search by Terms or Phrases
Search for keywords or expressions within the PDFs.

10. Quick Reference Guide — Pteryx

Step 1: Basic Requirements

Before installing, ensure your computer meets the minimum requirements:
• **Operating System**: Windows 10 or higher
• **Processor**: Intel Core i5 or equivalent
• **RAM**: 8 GB or more
• **Free disk space**: at least 2 GB
• **Microsoft Excel or equivalent installed** (for report handling)

Step 2: Download and Installation

1. Visit the official Pteryx website (www.pteryx.com.br).
2. Download the Pteryx installer.
3. Run the installer (Pteryx_installer.exe).
4. During installation, enter your e-mail when prompted.
5. The installer will download the necessary components and create the folder structure on your computer.
6. Wait for the installation to complete.

Step 3: Opening Pteryx for the First Time

1. After installation, open the program via the shortcut on the desktop or Start menu.
2. The software will automatically detect your system's language and configure the corresponding interface.
3. Register (if requested) or proceed according to the initial screen instructions.

Step 4: Preparing Your PDFs for Search

1. Gather the PDF files you wish to search.
2. Copy and paste these files into the 1_PDFs_to_Process folder, which was automatically created during installation.
   • You can find this folder in the directory where you installed Pteryx.
3. Alternatively, within the software, use the “Gather PDFs” function to select folders or drives to automatically copy the PDFs to the processing folder.

Step 5: Running Pre-processing

1. From the Pteryx main menu, click on Pre-processing.
2. Click on Start to begin pre-processing the PDFs.
3. The software will check your files, separating them into specific folders:
   • Corrupted PDFs → 4_PDFs_with_Errors
   • Duplicates → 5_PDFs_with_duplicates
   • PDFs requiring OCR → 2_PDFs_to_OCR
   • Ready-to-search PDFs → 3_PDFs_to_search

4. If there are files requiring OCR, the system will ask if you wish to perform Optical Character Recognition. Click Yes and wait for completion.

Step 6: Performing Searches in PDFs

1. After pre-processing, return to the main menu.
2. Click on Search to open the search module.
3. Choose between the three types of search:
   • Search by Species — to search for a specific species.
   • Search by Species List — to search for multiple species based on an Excel spreadsheet (Data.xlsx).
   • Search by Terms or Phrases — to search for keywords or phrases.
4. Configure the parameters according to the selected search type and click Start.
5. Wait for the reports to be generated.

Step 7: Viewing and Exporting Results

• Reports are automatically saved in the folder 6_PDFs_with_target_data.
• They will be organized in subfolders according to the species or term searched.
• Open the .xlsx files with Microsoft Excel or compatible software for detailed analysis.

Final Tips

• Always perform pre-processing before conducting searches.
• Keep your folders organized to avoid errors.
• Check the FAQ to resolve common issues.
• For technical support, contact **sac@pteryx.com.br.**
