**ReferenceUSA Data Scraper: User Guide**

**System Requirements**

* Python 3.7 or higher installed
* Chrome browser installed
* Internet connection
* Access to a valid library account with ReferenceUSA access

**Setup Instructions**

**For Windows Users:**

1. **Install Python** (if not already installed)
   * Download Python from [python.org](https://www.python.org/downloads/)
   * During installation, check the box that says "Add Python to PATH"
   * Complete the installation
2. **Install Required Packages**
   * Open Command Prompt (search for "cmd" in the Start menu)
   * Navigate to the folder containing the scraper files:
   * cd path\to\Data Axle Miller 3
   * Install required packages:
   * pip install -r requirements.txt
3. **Download ChromeDriver**
   * Check your Chrome version: open Chrome, click the three dots (⋮) in the top-right, select Help > About Google Chrome
   * Download the matching ChromeDriver version from [chromedriver.chromium.org](https://chromedriver.chromium.org/downloads)
   * Extract the downloaded zip file and place chromedriver.exe in the same folder as the scraper files (Data Axle Miller 3 folder)

**For macOS Users:**

1. **Install Python** (if not already installed)
   * Download Python from [python.org](https://www.python.org/downloads/)
   * Complete the installation
   * Alternatively, use Homebrew:
   * brew install python
2. **Install Required Packages**
   * Open Terminal (from Applications > Utilities > Terminal)
   * Navigate to the folder containing the scraper files:
   * cd path/to/Data\ Axle\ Miller\ 3
   * Install required packages:
   * pip3 install -r requirements.txt
3. **Download ChromeDriver**
   * Check your Chrome version: open Chrome, click Chrome in the menu bar, select About Google Chrome
   * Download the matching ChromeDriver version from [chromedriver.chromium.org](https://chromedriver.chromium.org/downloads)
   * Extract the downloaded zip file
   * Move chromedriver to the scraper folder:
   * mv ~/Downloads/chromedriver path/to/Data\ Axle\ Miller\ 3/
   * Make it executable:
   * chmod +x path/to/Data\ Axle\ Miller\ 3/chromedriver

**Running the Scraper**

1. **Configure the Scraper** (First-time setup)
   * Run the script once to create a configuration file:
     + Windows: Double-click on main.py or run python main.py in Command Prompt
     + macOS: In Terminal, run python3 main.py
   * The script will create a default configuration file and then exit
   * Open the config folder and edit referenceusa\_config.yaml with:
     + Your library card number and PIN
     + The state you want to search for
     + Any other preferences
2. **Run the Scraper**
   * Windows: Double-click on main.py or run python main.py in Command Prompt
   * macOS: In Terminal, run python3 main.py
   * The script will open Chrome and begin the process
   * Follow any on-screen prompts for manual actions when needed
3. **Monitor Progress**
   * The terminal window will show progress information
   * Downloaded files will appear in the downloads folder
   * The script will notify you of any issues that require your attention

**Pausing and Resuming**

1. **To Pause the Scraper**
   * Let the current batch finish downloading
   * Press Ctrl+C in the terminal/command prompt window
   * Or close the terminal/command prompt window
2. **To Resume**
   * Run the script again as described in "Run the Scraper" section
   * The script will automatically resume from where it left off
   * You can also specify a particular batch to start from when prompted

**Troubleshooting**

* **Chrome Opens but Nothing Happens**: Make sure your ChromeDriver version matches your Chrome browser version
* **Login Issues**: The script will prompt you to manually log in if automatic login fails
* **Downloads Not Working**: Check that you have sufficient disk space and permissions to write to the downloads folder
* **Script Crashes**: Check the error message in the terminal. Most common issues are related to timing or element detection, which can be resolved by restarting the script.

**Notes**

* The script handles captchas by pausing and asking you to solve them manually
* Each batch downloads approximately 250 records (10 pages)
* Your progress is automatically saved, so you can stop and resume at any time
* Files are saved with the naming format: StateName\_StartPageToEndPage\_DateTimeStamp.csv

**Getting Help**

If you encounter issues with the scraper, please:

1. Take a screenshot of any error messages
2. Note what you were trying to do
3. Contact technical support with this information

**File Locations in the ReferenceUSA Data Scraper**

**Downloaded Data Files**

The data files downloaded from ReferenceUSA are saved in:

Data Axle Miller 3/downloads/

* This is a subfolder named "downloads" inside your main "Data Axle Miller 3" folder
* Each downloaded CSV file will be saved here with names like:
  + New Mexico\_1to10\_090525\_1425.csv (State\_PageRange\_DateStamp\_TimeStamp.csv)
* These files contain the business data you've downloaded from ReferenceUSA

**Configuration File**

The configuration file is saved in:

Data Axle Miller 3/config/referenceusa\_config.yaml

* This YAML file contains all your settings:
  + Library credentials (card number & PIN)
  + State to search
  + Download batch size
  + File paths

**State/Progress Tracking File**

The script saves its progress in:

Data Axle Miller 3/config/reference\_usa\_state.json

* This JSON file tracks:
  + Which batches have been completed
  + URLs for search and results pages
  + Names of downloaded files
  + Total number of batches to download

**Important Notes:**

1. The script automatically creates these folders and files if they don't exist
2. You can change the download location by editing the download\_dir setting in the configuration file
3. If you move the script files to a new location, you should update the configuration file accordingly
4. Always check these folders after running the script to ensure files are being saved correctly
5. Keep these files intact if you want to pause and resume the scraping process

If you need to change where files are saved, edit the configuration file before running the scraper again.

**How to Stop the ReferenceUSA Scraper & Understanding State Files**

**How to Stop the Process**

**Graceful Stop (Recommended)**

1. **Wait for the current batch to finish downloading**
   * You'll see a message like "=== Processing batch X ===" followed by "Download complete"
   * After a batch completes, the script will save its progress automatically
2. **Press Ctrl+C in the terminal/command prompt window**
   * This sends a keyboard interrupt signal to stop the script
   * The script will exit cleanly after the current operation completes
   * You'll see "^C" or an interrupt message when you press Ctrl+C
3. **Close the terminal/command prompt window**
   * Simply closing the window will terminate the script
   * This works too, but Ctrl+C is slightly more reliable

**Emergency Stop**

If the script is unresponsive:

1. Press Ctrl+C multiple times
2. If that doesn't work, close the terminal/command prompt window
3. In extreme cases, you may need to close Chrome manually

**About the State File (reference\_usa\_state.json)**

**What is it?**

* This file tracks the script's progress and saves important information
* It's automatically created and managed by the script
* You don't need to create it manually

**What happens if it's missing?**

* If the file is missing when you start the script, a new one will be created automatically
* The script will start from the beginning (batch 1)
* Any previous progress tracking will be lost

**What's stored in it?**

The state file contains:

* Last batch processed
* Total number of batches
* List of completed batches
* URLs to return to when resuming
* List of downloaded files

**When is it updated?**

* After each successful batch download
* When search and results URLs are established
* When calculating total number of batches

**Importance for Stopping/Resuming**

* **DO NOT delete this file** if you want to resume where you left off
* The file is your "bookmark" that tells the script where to continue from
* If you want to start over completely, you can delete this file
* If you're changing search parameters, it's better to delete this file to avoid confusion

**In Summary:**

1. To stop: Press Ctrl+C or close the terminal window
2. No need to create reference\_usa\_state.json yourself - it's automatic
3. Keep the state file if you want to resume later
4. Delete the state file if you want to start fresh

The script is designed to handle interruptions gracefully, so don't worry about stopping it between batches - your progress is saved automatically.