Scholarly | Installation and Use

Before getting started, you'll need to install:

- R (version 4.1.2 or newer): https://cran.r-project.org/
- Git: https://git-scm.com/downloads (on a Mac it should be already installed or may need to be activated via dev tools)

Setup

Atlas Database

The scraped Google Scholar data will be securely stored in a cloud based MongoDB Atlas database for which only you have credentials.

1. Sign up
2. Create Organization
   a. Pick a Name
   b. Choose MongoDB Atlas
3. Create a Project
   a. You can pick any name you like, such as *Scholarly_universityName*
   b. No need to add additional users/permissions if a single person is using it
4. Click Build a Database
   a. Shared (FREE tier)
   b. Default settings are OK
   c. Click **Create Cluster**
5. Security Quickstart
   a. Create username (this is for the DB access), e.g. "dbaccess"
   b. Click **Generate secure password** (copy the password to a safe place)
   c. Create User
   d. Click **My Local Environment**→Add My Current IP Address
   e. Click **Cloud Environment**→Add the following IP addresses (via Add Entry):
      - 54.204.34.9
      - 54.204.36.75
      - 54.204.37.78
      - 34.203.76.245

        - 3.217.214.132
        - 34.197.152.155

    f. Click Finish & Close

6. From **Database Deployments**, find click **Connect** for the cluster we just created.
    a. Click **Connect your application**
    b. Click **Copy** button for mongo URL
    c. Replace `<password>` with the password generated at step 5.b

> **Retain this URL in a safe place, you will need it later**.

Scholarly Software - Windows
Note - Ensure **git** is installed before continuing!

Scholarly is the software that will collect the data and store it in the cloud Atlas DB. It should be installed on computer that will have a somewhat stable internet connection and be powered on during the data collection.

Click Windows Start Menu, then type **Command Prompt.** In the prompt, paste the following and press enter:

```
git clone https://github.com/DataAnalyticsinStudentHands/Scholarly.git
```

Wait a few moments for download; messages in the command prompt will confirm successful download and unpacking.

In the same command prompt, run the following command:

```
setx PROJECT_DIR C:%HOMEPATH%\Scholarly
```

Run the following command, except replace the text "pasteMongoAtlasURLHere" with the URL that you copied at the end of the "Atlas Database" setup section.

```
setx MONGOURL pasteMongoAtlasURLHere
```

Run the following commands one by one:

```
cd %PROJECT_DIR%
setx R "C:\Program Files\R\R-4.1.2\bin\R.exe"
setx Rscript "C:\Program Files\R\R-4.1.2\bin\Rscript.exe"
```

Paste the following command in as one line:

```
reg add "HKCU\Software\Microsoft\Command Processor" /v Autorun /d "doskey
/macrofile=\"%PROJECT_DIR%\renv\scrape_macro.doskey\"" /f
```

Lastly, paste the following:

```
R
renv::restore()
```

The last command may take a minute or two to finish. You may be prompted to "select a download mirror" - enter `78` *if prompted, or if a popup appears, select one of the USA Mirrors such as* `USA (TX 1)` *. It will continue to install for several minutes, you will know it is finished when the prompt (*`>`*) reappears*.

After the last command finishes successfully, you can close the command prompt.

Performing Updates

Scholarly is periodically updated. To update to the latest version, use Windows Explorer to navigate to the Scholarly directory. Right click on an empty area in the window to open the context menu, and click `Git Bash here` then type the command:

```
git pull
```

Note: If Git Bash is not found on the context menu, you can find it from the Start menu. Use the `cd` *command and the filepath to navigate to the Scholarly directory.*

Scholarly Software - Mac

Note - Ensure **git** is installed before continuing!

Scholarly is the software that will collect the data and store it in the cloud Atlas DB. It should be installed on computer that will have a somewhat stable internet connection and be powered on during the data collection.

Open the Terminal app on your Mac, .e.g. by searching via ➝ "Command+Option" key and typing "terminal". In the terminal run the following command.

```
git clone https://github.com/DataAnalyticsinStudentHands/Scholarly.git
```

Wait a few moments for download; messages in the command prompt will confirm successful download and unpacking. If terminal prompts to install Xcode developer tools, accept installation and follow onscreen instructions.

In the same terminal, run the following command:

```
vi ~/.zprofile
```

use the ➝ i key to enable editing mode and add the following line

```
export PROJECT_DIR=~/Scholarly
```

Then add the following lines, except replace the text "pasteMongoAtlasURLHere" with the URL that you copied at the end of the "Atlas Database" setup section. Make sure you leave the double quotation marks around it.

```
export MONGOURL="pasteMongoAtlasURLHere"

alias scrape='
cd $PROJECT_DIR
echo "Starting scraper!"
Rscript scrape_control.R
cd -
'
```

Close the vi editor by using the following keys ➝ Esc, :wq. After pressing Esc you should be able to enter the colon at the bottom and you can enter wq.

Open a new terminal window and run the following command:

```
cd $PROJECT_DIR
```

Lastly, open an R terminal by running:

```
R
```

Once the R terminal is open run the following

```
renv::restore()
```

The last command may take a minute or two to finish. You may be prompted to "select a download mirror" - enter `78` if prompted, or if a popup appears, select one of the USA Mirrors such as `USA (TX 1)` . It will continue to install for several minutes, you will know it is finished when the prompt (`>`) reappears.

Performing Updates

Scholarly is periodically updated. To update to the latest version, open Terminal and run the following command.

Note: Ensure that you are in the same directory in which you ran the `git clone` command during installation.

```
cd Scholarly
git pull
```

Use

Introduction

In general there are three phases associated with this workflow:

1. Specifying "target" scholar profiles (after a pre-flight check of the input data has been completed)
2. Data collection
3. Data analysis/export

Phase 2 is the most time intensive, while phases 1 and 3 can generally be completed in a few minutes each using the Scholarly web app which can be accessed via a browser.

There will be three **collections** created in the MongoDB Atlas Database that you created earlier:

1. **scholars** - Links your file (ID number and Google profile link) with data collection, and monitors progress
2. **googleProfiles** - Contains Google profile-level information about scholars

3. **googlePublications** - Contains publication-level information about publications linked to scholar profiles

Preflight Data Check

This process requires data to be provided in a specified, standardized format using the "Coding Sheet" in XLSX format. The preflight check helps catch common errors from the data entry phase, such as duplicated IDs or mismatched IDs.

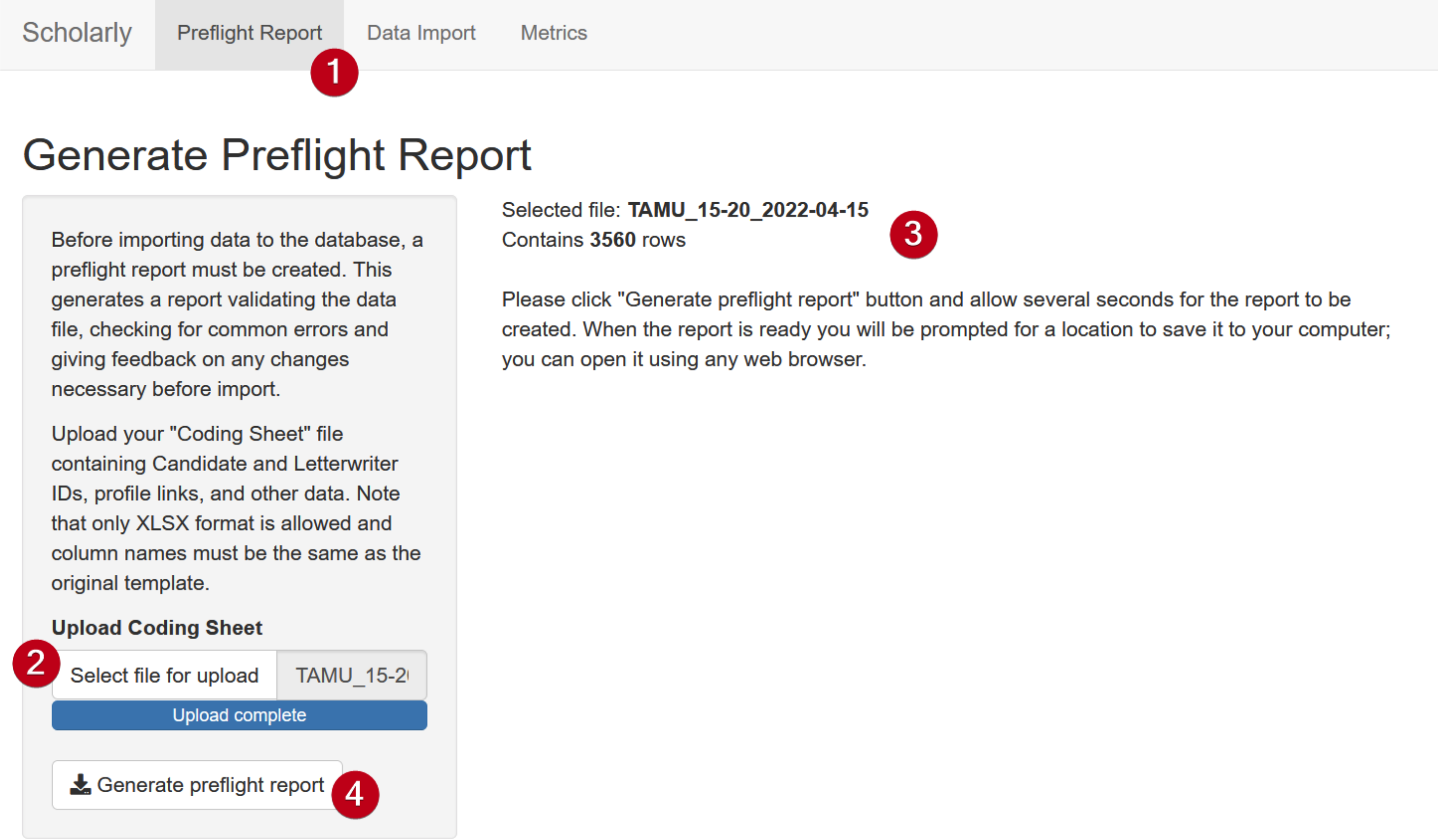
1. Navigate to the Scholarly Web App - Preflight Report tab
2. Click **Select File for Upload** and navigate to and select your "Coding Sheet" file

    > Note that only required columns (IDs and GS links) are processed by the web app. No data from your file is retained, communicated, analyzed, or stored.

3. Once the upload is complete, verify on the sidebar that the file name and number of rows matches your expectation
4. Click **Generate preflight report** and select the location on your computer where you would like to save the report file
5. Review the preflight report by opening it in any web browser such as Chrome, Safari, Edge, Firefox, etc.
6. Make any amendments to your file as necessary

    > Note that any **error** identified by the report **must** be addressed prior to proceeding to next steps.

    > Warnings identified by the preflight report are informational, if you identify that the warning is acceptable, you may proceed without amending the file.



Generate Preflight Report

Before importing data to the database, a preflight report must be created. This generates a report validating the data file, checking for common errors and giving feedback on any changes necessary before import.

Upload your "Coding Sheet" file containing Candidate and Letterwriter IDs, profile links, and other data. Note that only XLSX format is allowed and column names must be the same as the original template.

Upload Coding Sheet

Select file for upload | TAMU_15-2

Upload complete

Generate preflight report

Selected file: **TAMU_15-20_2022-04-15**
Contains **3560** rows

Please click "Generate preflight report" button and allow several seconds for the report to be created. When the report is ready you will be prompted for a location to save it to your computer; you can open it using any web browser.

Populating the database

> Ensure there are **no errors** in your **preflight report** before proceeding with this section

Before initiating data collection, you will need to specify the "targets," in the form of Google profile links that were recorded in your "Coding Sheet."

 These steps can be done on any computer, it is not necessary to use the same machine that will be purposed for data collection.

1. Navigate to the Scholarly Web App - Data Import tab
2. Paste your **MongoURL** that you created in **Setup - Atlas Database** step into the `MongoDB Atlas Connection` field. The sidebar will update to show your database and username (and number of existing records, for example, if this is not the first import).

   > An error will display if the URL is not correct, in which case you must verify proper setup - especially that the URL has been correctly configured with username and password, and the specified IP addresses were whitelisted in the Security panel.

1. Click **Select File for Upload** and navigate to and select your "Coding Sheet" file

    Note that only required columns are processed by the web app. No data from your file is retained, analyzed, or stored. Required data from your file will be communicated only to your Atlas Database.

2. Click the **Validate Data & Connection** button and ensure there are no errors displayed
3. Click the **Import Data to MongoDB** to initiate the upload.

After the upload is finished, you are ready to begin data collection!
You can close or navigate away from the app.

Data Collection

Once the initial database import has completed, the data collection computer(s) can begin collection process.

1. **WINDOWS**: Click Windows Start Menu, then type **Command Prompt.**

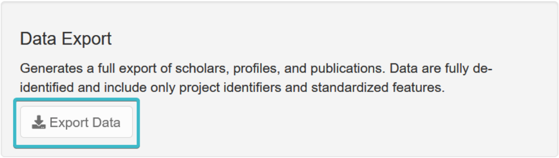
   MAC: Open **Terminal**
2. In the prompt, paste the following and press enter:

   ```
   scrape
   ```
3. Progress should be displayed on the command prompt window. Do not close the command prompt window or put the computer to sleep.
4. Data collection will automatically halt when finished and display a completion message.

The same command can be run on multiple computers (simultaneously if desired), or used to restart the process in the case of interruption (such as loss of internet connectivity).

Exporting Data

1. Navigate to the Scholarly Web App - Metrics tab
2. Paste MongoDB Atlas Connection URL into indicated field
3. Click the Data Export button

   Data Export

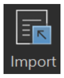
   Generates a full export of scholars, profiles, and publications. Data are fully de-identified and include only project identifiers and standardized features.

4. When the download is ready, save it in the desired location on your computer. The download is for a zip file, which contains de-identified productivity information. Copy the confirmation text on the right side of the "Connect to Atlas Database" panel which states "Contains [#] scholar records & [#] publications"
5. Email or use a file-sharing service to transmit the file to the recipient party. Paste the text that you copied in the last time to allow recipient to confirm the correct number of documents transmitted.

Importing Transferred Data

Upon receipt of the data export from a partner, the data can be directly imported to MongoDB to be folded into analytic pipelines. This tutorial uses Studio3T as the import manager, but any other MongoDB interface should have similar functionality.

Import:

1. Unzip data export folder.
2. There are three `json` files, one for each collection (scholars, googleProfiles, googlePublications)
3. **Precheck**
    a. Calculate "completeness" for records to be imported (script)
    b. Doesn't need to be perfect, but if many are incomplete, discuss whether partner should continue scrape then re-export prior to import
4. Connect to the appropriate database in Studio3T
5. Right click/open context menu on the destination collection and select `Import data...` or click the Import button 

6. Leave at default option - JSON and click `Configure`
7. Click `+Add Source`
8. Select the `json` file that matches the collection. Check the preview to ensure there is no error.
9. Click `▶ Run` to load the data.
10. Repeat for the other two collections.