# Instructions on the use of the Scraper\_GUI

* + - 1. Before you start, please make sure you have internet and that your computer does not go to sleep while this scrapes. If it does go to sleep, chrome will also disconnect, and the scraper may crash or face an infinite loop (which you must force quit out of by right clicking on the icon in the hotbar and selecting “quit”). With most crashes (except this) the scraper will attempt to create a partial file which will start off where you left off when you start the same search again. Check the terminal for crash reports and contact [ericchantland@gmail.com](mailto:ericchantland@gmail.com) if you are faced with a crash that does not save your file and has not been already described. Additionally, make sure you are either logged into a valid university VPN or have an account which has access to eHRAF searches. For limitations of the Scarper GUI (what it cannot do) please see Limitations: below
      2. Graphical user interface, text, application

         Description automatically generatedDouble click on the GUI icon wherever it was downloaded to.

Text

Description automatically generatedThis should produce a Terminal which has valuable information not shown within the GUI. If the GUI ever crashes, the error code will be located there. Additionally, the terminal will tell you if you have a partial save file present.

Note that it takes a few seconds before the actual GUI starts.

Graphical user interface, application, PowerPoint

Description automatically generatedThen you should see the actual Scraper GUI:

* + - 1. Graphical user interface, text, application

         Description automatically generatedEnter your **Name** within the name line. Then, when you are finished, click on the **Search** tab
      2. Graphical user interface, application, chat or text message

         Description automatically generatedNow you should be on the Search tab. There are two ways of searching: Through a URL or through an Advanced Search. You will usually use the Advanced Search feature anyway so let’s focus on that. Besides, using an URL requires that we already have a URL… which you might not have yet…
      3. Starting with the Advanced Search, you can put in any number of **cultures**, **subjects**, or **keywords** separated by a comma. Here is an example, say we wanted to search in the culture of Azande, for the subjects/OCMs 751, 77, and Diet and the lexical keywords Apple and Pear. Here is how we would input it:

Graphical user interface, application

Description automatically generatedNotice that you can put numbers and words into the subject box. So long as they are valid OCMs, it will work! Just make sure you have a comma separating each query input (search term).

* + - 1. Graphical user interface, text, application, chat or text message

         Description automatically generatedWhen you are ready, push the button “SUBMIT”
      2. The program will open up a chrome web browser and catalog how many passages it is about to scrape as well as the URL’s for the cultures. However, the above query inputs we submitted were actually **TOO SPECIFIC** and resulted in no search results (passages) found.

Graphical user interface, application

Description automatically generatedIf these query inputs were what you originally wanted, then just move on and catalog that the search gave no results. However, there are other things you can also do to affect the search.

* + - 1. *Graphical user interface, application

         Description automatically generated*You can change how the query inputs are searched for by pressing the blue buttons. These affect what the passage must contain in order to be scraped. For example, the GUI pictured below would try to scrape passages which are **NOT** in the Azande culture, have **ANY** of the following OCMs (750; 77; diet/262), **AND** have **ALL** of the following keywords (Apple or Pear). Additionally, only passages part of the cultural level **PSF** will be selected. Now try submitting again. Note that, unless otherwise told, you likely will never change a checked box in the “Cultural Level” filter
      2. Graphical user interface, text, application

         Description automatically generatedIf there are passages that fit our search query (from the query inputs we provided) we will get info on non-valid inputs (77 is not a real OCM), how may passages will be scraped, and the generated URL that is in use. (Note that if you used the “URL input” search, the URL in use is whatever you pasted in)
      3. Graphical user interface, text, application, chat or text message

         Description automatically generatedIf you like the search about to happen, press the CONTINUE button and the scraping will start.
      4. Graphical user interface, text

         Description automatically generated When the Scraping is finished, the chrome page will close and the GUI will tell you it is finished. You should also check the terminal for more information but this is optional.
      5. If the scraper should crash or finish to completion, a partial file may be created which can be used to start where you left off. To do so, just use the same inputs you used before in the Advanced Search OR copy and paste the URL found in the data file created. Sometimes, an unforeseen crash may occur which does not initiate the partial save feature. If this happens, it may be because the website has changed since last version implementation or another rare problem occurred. Consider changing the Partial save frequency detailed in the **OPTIONS** below.

### **Scraping via a URL:**

### Graphical user interface, text, application Description automatically generatedScraping just using a URL can be good if you already have done the same query search before this. You would simply take any URL generated from the “URL in use” box (Copy cmd+c and Paste cmd+v) …

A screenshot of a computer screen

Description automatically generated with medium confidenceTable

Description automatically generated with low confidenceOr an excel data file…

*Graphical user interface, application

Description automatically generated* and input it into the “eHRAF URL” line in the GUI!

### Graphical user interface, text, application, chat or text message Description automatically generatedThere is a second way of getting the URL link, through the eHRAF database yourself. You can find Advanced Searches [here](https://ehrafworldcultures.yale.edu/search/advanced) or navigate through eHRAF online. The Advanced Search has the exact same parameters as the GUI except that you cannot enter numbers for OCMs (if you do, it will ask you to select from the drop down what you mean).

Graphical user interface, text

Description automatically generatedYou may also add a second line of queries by pushing the + button. This second line is done in addition to the subjects and keywords added above. It still uses the cultures you added in the first line. This is a great way to have even more specified and refined searches but is probably not something you will often or ever use.

* + - 1. Graphical user interface, text, application, email

         Description automatically generatedClick Search when you are ready. This will give you the results page.
      2. To get filters to specify your search (like getting only PSF cultures). Click on “**Show filters**” and a filter column will be shown. If you wanted to only get PSF cultures, you might choose “**Cultural Level Samples**” and click on “**PSF**”

Graphical user interface, text, application, email

Description automatically generated

* + - 1. When you are ready, copy the entire URL on the eHRAF page and paste it in the GUI so that it can automatically scrape the webpage you pasted in.

Graphical user interface, text, application

Description automatically generated*Graphical user interface, application

Description automatically generated*

* + - 1. Regardless of how you got the URL, push SUBMIT to start the Scraping

### **Options:**

### Graphical user interface, text, application Description automatically generatedOptionally, you can also click the “Options” tab for parameters which will affect the scraper but not the passages it scrapes (meaning that whatever you pick will not affect the number of passages you get). Shown below.

### You can choose whether the browser is visible while it scrapes. The default is yes.

### Graphical user interface, application, website Description automatically generatedFor “Use partial files if present?” the scraper will automatically save a partial excel file if it encounters MOST problems that it cannot handle (like the internet going out). Selecting “NO” will make the scraper start from the beginning even if there is already a partial file with the same search parameters present. Therefore, it will overwrite the old file.

### For “Display number of passages per culture”, selecting “YES” will display the count of passages per culture within the GUI terminal. The order of the cultures can be alphabetical (by selecting “culture”) or by “count”. This will also affect what order the cultures are scraped so it is nice way to know your progress.

### A screenshot of a computer Description automatically generated with medium confidenceFor “Save separate culture files”, Selecting “YES” will output individual cultural scrapings in addition to the altogether dataset.

### For “Document Save Iteration” the scraper will attempt to save itself every x number of passages that it successfully scraped. The default is 5,000 but you can change it to how you see fit or select “None” if you don’t want it to be saved at all. The lower the number the more times it will be saved. Although saving is rather quick, it still takes time so beware. Note that partial saves only occur after a culture is finished being scraped so if you want it to save after each culture scraped, set the save iteration to 1

### **Limitations:**

### Note that most limitations here can be circumvented by obtaining a URL from eHRAF’s [Advanced Search](https://ehrafworldcultures.yale.edu/search/advanced) feature. Any URL generated can be scraped using the GUI’s URL search line.

* Currently (although this will likely be added later), the GUI cannot create an extra subject/text clause like that in eHRAF’s Advanced Search (detailed in paragraph ‎(2) above)
* Filters excluding “Culture Level Samples” are not implemented such as “Subsistence Types”, “Regions”, “Published Date”, “Document Level Samples”, “Document Types” and “Series”