

## Infectious Diseases - Data Cleanup Handover Document

**Group Members:** Ellie Jensen, Andrew Clements, Collin Bowers, Ziyun Ma

**Detailed instructions for someone starting the project where you are leaving it**

Brief Background Information:

The Carter Center is dedicated to advancing human rights and alleviating suffering and one of their focuses is on minimizing infectious diseases. To do this, they collect data from geographic units within each country, but the data is often given in the native languages of these countries, leading to conflicting translations. A similar project was previously undertaken for the country of Sudan, and this semester the Carter Center is looking for help with their Ethiopian data. The primary challenge is due to the lack of resources for Amharic and the absence of a standard English transliteration. To address this, a tool, along with an additional GUI, was developed to transliterate a list of region, zone, and woreda inputs using a standardized list of target mappings. The following instructions provide guidance on running the Google Colab tool and Anvil GUI (Note: links to the code can be found at the bottom of this document).

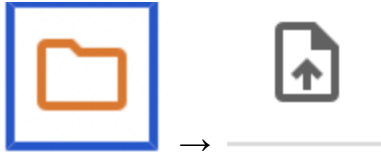
Instructions for running our project (Google Colab Ethiopian Tool): The Ethiopian Tool allows you to generate standardized regions, zones, and woredas directly, without using the Anvil GUI. Follow these steps to run the tool:

### 1. Specify Input and Output Details:

- a. In the "Inputs and Outputs" section of the code, enter the name of your input file in the 'input\_file' box (an example one is linked at the bottom of this document). This file should have regions in the first column, zones in the second, and woredas in the third.
- b. Enter the name of your master/target mapping file as the 'target\_mapping\_file' (an example one is linked at the bottom of the document). This file must be in Excel format, with regions, zones, and woredas listed in the first column, separated by underscores (region\_zone\_woreda).
- c. Indicate the specific sheet within your input file that you wish to transliterate.
- d. Provide a name for the output file you want to create.
- e. Set the confidence levels for the algorithm in the low\_confidence\_score and xmed\_confidence\_scorefields. (We recommend using 0.3 for low confidence and 0.4 for medium confidence. Higher values increase the likelihood of mappings being flagged for review because they require more precise matches.)

### 2. Upload Files:

- a. Ensure the uploaded input and mapping files match the names specified in the input boxes.
- b. To upload files, click the file icon in the Colab sidebar, then click the file upload icon and select your files.



3. **Run the Program:**
  - a. Navigate to Runtime in the menu and select Run All to run the program.
4. **Download and Review the Output:**
  - a. Once the program finishes running, the output file will appear in the file section. Double-click to download it.
  - b. When opening the file in Excel, you might get alerts about changes. For the first alert, choose "Yes," and for the second alert, choose "Delete."
5. **Examine the Results:**
  - a. The new column of mappings will appear as the first column in the file.
  - b. Any mappings that are recommended for review will be highlighted (anything highlighted in red is more likely to be incorrect than anything highlighted in orange)

Instructions for running our project (Anvil GUI): We made this upon request of the client to make this tool easily useable (without having to directly deal with code):

1. **Required Files:**
  - a. **Target/Master Mapping File:** This file contains the standardized naming of regions, zones, and wordas in the format region\_zone\_worda (an example one is linked at the bottom of this document).
  - b. **Input File:** This file contains the data to be standardized, with regions in the first column, zones in the second column, and wordas in the third column (an example one is linked at the bottom of this document).
2. **Steps to Use the Tool:**
  - a. Open the Anvil website link.
  - b. Upload the Target Mapping File and Input File. If no target mapping file is uploaded, the tool will use the default file (linked at the bottom).
  - c. Specify the name of the sheet in the input file you want to process in the user input field titled sheetname.
  - d. Enter a name for the output file. If no name is provided, the default name will be used.
  - e. Set the confidence levels for the algorithm:
    - i. Low Confidence Score: Typically set to 0.3.
    - ii. Medium Confidence Score: Typically set to 0.4.  
(Higher confidence scores result in more mappings being flagged for review because they require more precise matches.)
3. **Generate the Output:**
  - a. Click "Generate" to run the program.
  - b. A pop-up will appear, allowing you to download the output Excel file to your computer (For the first alert, choose "Yes," and for the second alert, choose "Delete")
4. **Review the Output:**

- a. The file will contain a new column with standardized mappings.
- b. Any mappings requiring review will be highlighted for your attention.

### **A description of the major project goals, and which of those you worked on this semester**

Project Goals from our original project document were:

- **(Primary Goal)** Have a working tool that uses hardcoded mappings of the Ethiopian woredas (district) names to output the standardized spellings of a list of input into an excel file and marks cases that should be reviewed by the user.
- **(Secondary Goal 1)** Have a general working tool so that the user can input any standardized mapping file for any language/country they are working with and get an output of the standardized spellings in an excel file that marks cases that should be reviewed by the user.
- **(Secondary Goal 2)** Have a working tool that takes into consideration changes in district boundaries and is able to map previous boundaries to the current boundaries, with the cases that should be reviewed by the user marked in the excel file.

Additional goals added during the semester:

- **Front-End GUI:** At the client's request, we prioritized developing a GUI using Anvil to make it easier for non-technical users to successfully utilize the tool.
- **Refactoring Code:** We refactored the Google Colab code to streamline its structure and make it easier to create the general solution for any country in the future.

Progress on goals made during the semester:

- **Google Colab Ethiopian tool:** The primary goal was completed with this tool. We developed and refined the tool through multiple iterations and tested it extensively to ensure accuracy. In addition to what our original primary goal was, the tool improves upon the initial hardcoded mappings approach by allowing users to upload a standardized target/master mapping file. This flexibility means that users can quickly update the tool to reflect changes in Ethiopian regions, zones, and woredas by just modifying the uploaded file. The output is an Excel file containing standardized spellings, with ambiguous cases flagged for user review.
- **Anvil GUI:** Based on discussions with the client, we prioritized building a front-end GUI for the tool before generalizing it for use with other languages or countries. The GUI is built using Anvil. It mirrors the functionality of the Google Colab tool while offering a more intuitive interface, eliminating the need for users to interact with the underlying code.
- **Generalized tool (can be used for any country):** To make it easier to create a general tool, we refactored the Google Colab code into modular functions, making it easier to adapt for future applications. The refactored code is organized so that all the code is encapsulated in functions and those functions are called from the implementation block. This structure sets up the tool to be adapted to work with various languages and countries. The generalized tool is still a work in progress (**Not Completed**), but the ideas

and existing code are in the "Copy (for general solution) of Code\_Locality\_Standardizer\_Final.ipynb" file in Google Colab (linked below).

### **A clear idea of the next few steps for continuing the project**

- **Finish generalized tool (can be used for any country):** The next step would be to finish the general tool that can be used for any country by adapting the Ethiopian tool for varying geographic structures beyond the Ethiopian context. This task was started but not completed. We explored strategies to accommodate any number of geographic layers, such as regions, zones, woredas, districts, or additional hierarchical levels. The generalized tool will allow users to simply enter their input file and target mapping for any data set they have (the files still need to follow the same structure every time). The tool will then produce outputs that adhere to the same standardized format, highlighting unsure cases for review.
- **District Boundary Updates and Mapping:** After that is complete, the client has expressed interest in incorporating functionality that accounts for changes in district boundaries that are reported online. This would involve pulling updated boundary data directly from online resources. The tool would then map previous district boundaries to the new ones, which would ensure that the tool is taking the current boundaries into account. Just like the current tool, the updated version will flag cases that require user review by highlighting them.

### **An update on any interaction or contact with the client - specifying who the client is and how to reach them**

The primary clients are **Jenna Coalson** (Jenna.Coalson@cartercenter.org) and **Emily Griswold** ([emily.griswold@cartercenter.org](mailto:emily.griswold@cartercenter.org)) because they will be the primary users of the Ethiopian Transliteration Tool. We mainly communicated with them over email and on zoom meetings. They both respond to emails in a timely manner (usually on the same day), but they do travel to the countries they work with, so there could be delays in response while they are traveling.

We also are presenting to the Carter Center during their monthly lunch and learn on December 11th, and our main point of contact for that was **Ursula Kajani** (Ursula.Kajani@cartercenter.org) over email.

### **Demo videos of what state your project is in right now -**

#### **\* Demo - Google Colab Ethiopian tool:**

 Demo\_Colab\_Solution\_12:4:2024.mov

#### **\* Demo - Anvil GUI:**

### **Code walkthrough videos -**

#### **\* Code Walkthrough - Google Colab Ethiopian tool:**

 Code\_Walkthrough\_Colab\_Ethiopian\_Solution\_12:4:2024.mov

#### **\* Code Walkthrough - Google Colab general tool (not completed):**

 [Code\\_Walkthrough\\_Colab\\_General\\_Solution\\_12:4:2024.mov](#)

**\* Code Walkthrough - Anvil GUI:**

**Link to Google Colab Ethiopian Tool Code:**

 [Code\\_Ethiopia\\_Locality\\_Standardizer\\_Final.ipynb](#)


**Link to Anvil Code:**

 [FINAL Anvil\\_Code\\_Ethiopia\\_Locality\\_Standardizer\\_Final.ipynb](#)


**Link to the Generalized Tool Code (started but not complete):**

 [Copy \(for general solution\) of Code\\_Locality\\_Standardizer\\_Final.ipynb](#)

**Link to Input File Used in Demo:**

 [Sample data file\\_river prospection copy.xlsx](#)

**Link to Mapping File Used in Demo:**

 [2024\\_ETH\\_RBLF\\_Admin\\_Names\\_Compiled copy.xlsx](#)