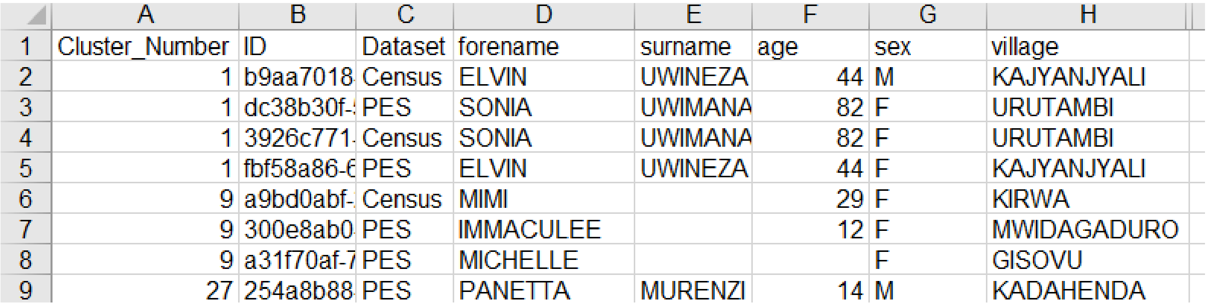
# CROW Cluster Version; Instructions For Early Release

Please note, all exemplar data displayed in this document is fake, synthetic data.

## Input File Requirements

The requirements for the files that you wish to clerically review are as follows;

- The files must be CSV files.
- They must be 'long' files. There must be one row per record, containing identifiable variables that are to be presented to clerical matchers for review. Clusters will contain multiple rows. See below for an example
- There must be a column that represents cluster ID. These ID's can be any random string, so long as they are unique to each cluster.
- There must be a record ID column. This must be unique for each record.



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Cluster_Number | ID | Dataset | forename | surname | age | sex | village |
| 2 | 1 | b9aa7018 | Census | ELVIN | UWINEZA | 44 | M | KAJYANJYALI |
| 3 | 1 | dc38b30f- | PES | SONIA | UWIMANA | 82 | F | URUTAMBI |
| 4 | 1 | 3926c771- | Census | SONIA | UWIMANA | 82 | F | URUTAMBI |
| 5 | 1 | fbf58a86-6 | PES | ELVIN | UWINEZA | 44 | F | KAJYANJYALI |
| 6 | 9 | a9bd0abf- | Census | MIMI | | 29 | F | KIRWA |
| 7 | 9 | 300e8ab0 | PES | IMMACULEE | | 12 | F | MWIDAGADURO |
| 8 | 9 | a31f70af-7 | PES | MICHELLE | | | F | GISOVU |
| 9 | 27 | 254a8b88 | PES | PANETTA | MURENZI | 14 | M | KADAHENDA |

Example 'long' file data format to be input to CROW

## Setting Up The Config File (for project leads)

In order to adapt the cluster version of CROW to your data, you will need to adapt the config_clusters file to meet the needs of your project. Instructions for doing this are self-contained within the config file, however, below are the key points.

**[column_headers_and_order]**
This is where you can specify what text you want to appear as the header for each row. This can be independent of the column headers in your CSV. You only need to create titles for the variables you want to display.

**[column_file_info_and_order]**
This is where you specify how CROW reads the variables in your CSV. Variable titles in this field MUST be the same as in your CSV. You only need to create titles for the variables you want to display.

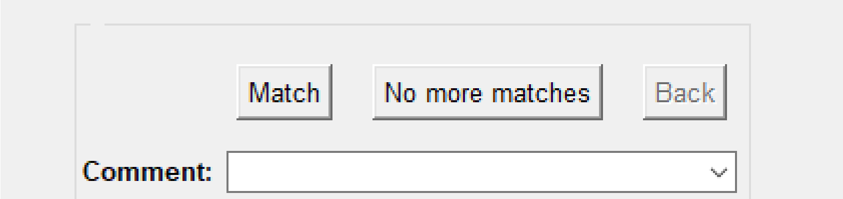
**[cluster_id]**
Specify which column is your cluster id column (MUST be the same as in your CSV).

**[record_id]**
Specify which column is your record id column (MUST be the same as in your CSV).

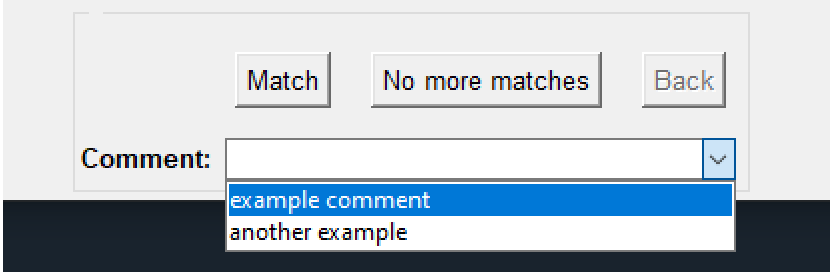
### [custom_settings]

There are a few optional custom settings that can be personalised for matching projects. These include:

Commentbox: when set to 1, a comment box is displayed as shown below that allows clerical matchers to enter comments that will be appended to the relevant clusters in the matched file. If set to 0, this is not displayed.

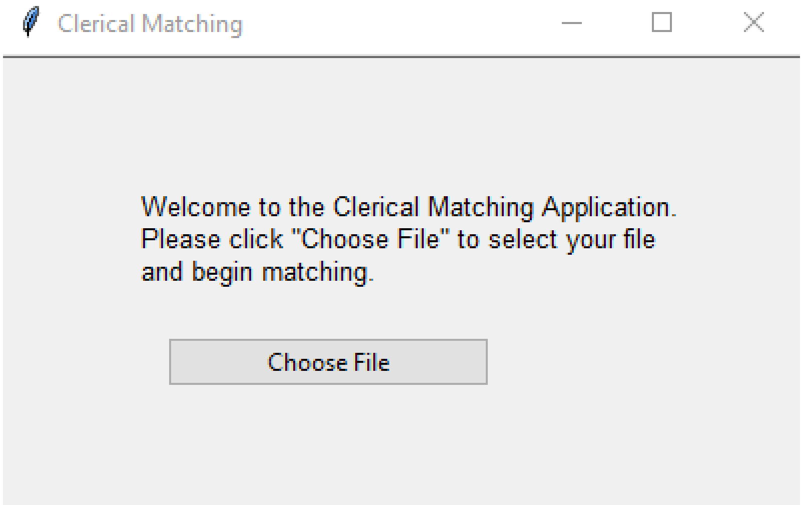


comment_values: shows example dropdown options for the comment box that can be selected by clerical matchers. If left blank, no options are shown.



## Using The Cluster Review Version

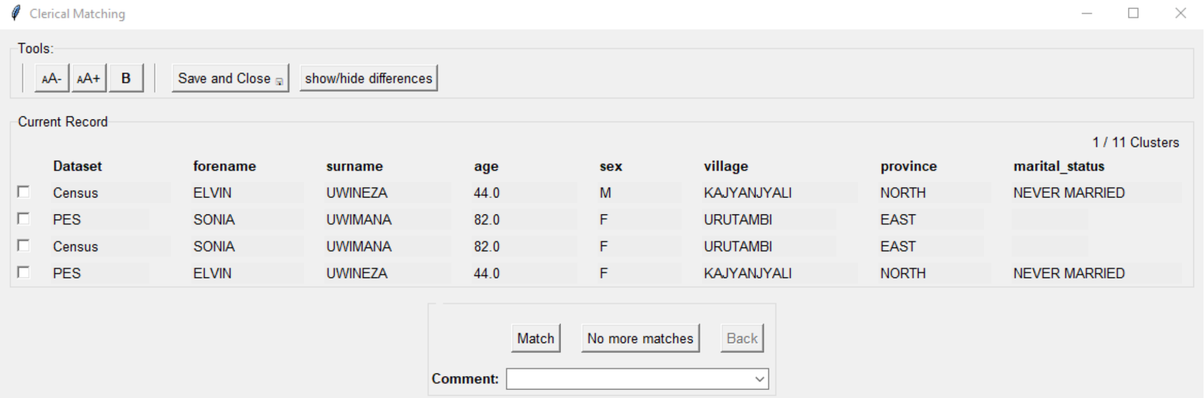
To use the Cluster Version of CROW; users can simply click play in their given python editor. The below window will then pop-up.



Welcome to the Clerical Matching Application.
Please click "Choose File" to select your file
and begin matching.

Clerical matchers can then choose the file they wish to match.

After selecting a file to match, the below window will pop-up.



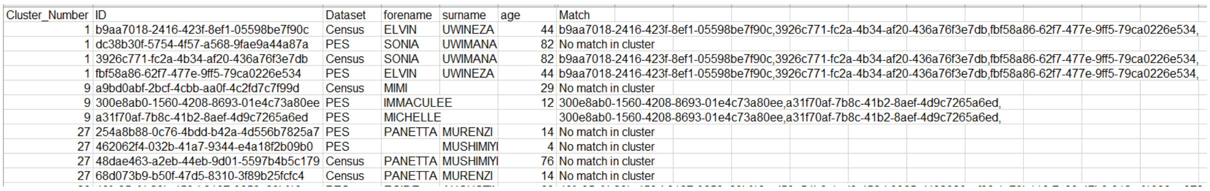
| | Dataset | forename | surname | age | sex | village | province | marital_status |
|---|---|---|---|---|---|---|---|---|
| □ | Census | ELVIN | UWINEZA | 44.0 | M | KAJYANJYALI | NORTH | NEVER MARRIED |
| □ | PES | SONIA | UWIMANA | 82.0 | F | URUTAMBI | EAST | |
| □ | Census | SONIA | UWIMANA | 82.0 | F | URUTAMBI | EAST | |
| □ | PES | ELVIN | UWINEZA | 44.0 | F | KAJYANJYALI | NORTH | NEVER MARRIED |

Users can then use the checkboxes to highlight records they wish to match, followed by clicking the ‘Match’ button. There can be multiple matches in a given cluster. Once all matches are exhausted, or the “No more matches” button is clicked, CROW will move on to the next cluster for resolution.

The back button can be used to return to the previous decision. Pressed the first time it will reset the current cluster (if some decisions have already been made in that cluster). Pressed again it will take the user back to the previous cluster.

## The output

The output file will look something like this:



| Cluster_Number | ID | Dataset | forename | surname | age | Match |
|---|---|---|---|---|---|---|
| 1 | b9aa7018-2416-423f-8ef1-05598be7f90c | Census | ELVIN | UWINEZA | 44 | b9aa7018-2416-423f-8ef1-05598be7f90c,3926c771-fc2a-4b34-af20-436a76f3e7db,fbf58a86-62f7-477e-9ff5-79ca0226e534, |
| 1 | dc38b30f-5754-4f57-a568-9fae9a44a87a | PES | SONIA | UWIMANA | 82 | No match in cluster |
| 1 | 3926c771-fc2a-4b34-af20-436a76f3e7db | Census | SONIA | UWIMANA | 82 | b9aa7018-2416-423f-8ef1-05598be7f90c,3926c771-fc2a-4b34-af20-436a76f3e7db,fbf58a86-62f7-477e-9ff5-79ca0226e534, |
| 1 | fbf58a86-62f7-477e-9ff5-79ca0226e534 | PES | ELVIN | UWINEZA | 44 | b9aa7018-2416-423f-8ef1-05598be7f90c,3926c771-fc2a-4b34-af20-436a76f3e7db,fbf58a86-62f7-477e-9ff5-79ca0226e534, |
| 9 | a9bd0abf-2bcf-4cbb-aa0f-4c2fd7c7f99d | Census | MIMI | | 29 | No match in cluster |
| 9 | 300e8ab0-1560-4208-8693-01e4c73a80ee | PES | IMMACULEE | | 12 | 300e8ab0-1560-4208-8693-01e4c73a80ee,a31f70af-7b8c-41b2-8aef-4d9c7265a6ed, |
| 9 | a31f70af-7b8c-41b2-8aef-4d9c7265a6ed | PES | MICHELLE | | | 300e8ab0-1560-4208-8693-01e4c73a80ee,a31f70af-7b8c-41b2-8aef-4d9c7265a6ed, |
| 27 | 254a8b88-0c76-4bdd-b42a-4d556b7825a7 | PES | PANETTA | MURENZI | 14 | No match in cluster |
| 27 | 462062f4-032b-41a7-9344-e4a18f2b09b0 | PES | | MUSHIMIYI | 4 | No match in cluster |
| 27 | 48dae463-a2eb-44eb-9d01-5597b4b5c179 | Census | PANETTA | MUSHIMIYI | 76 | No match in cluster |
| 27 | 68d073b9-b50f-47d5-8310-3f89b25fcfc4 | Census | PANETTA | MURENZI | 14 | No match in cluster |

A match column is appended to each row. If the given record has a match, the
record

A match column is appended to each row. If the given record has a match, the record for that record and all the matches will be printed in the ‘Match’ column separated by a comma.

If there are no matches to a given record ‘No Matches In Cluster’ is printed in the ‘Match’ column.