

# CROW Cluster Version Instructions

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	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	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).

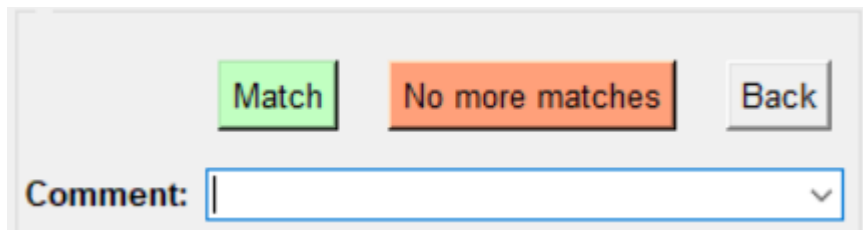
### [record\_id]

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

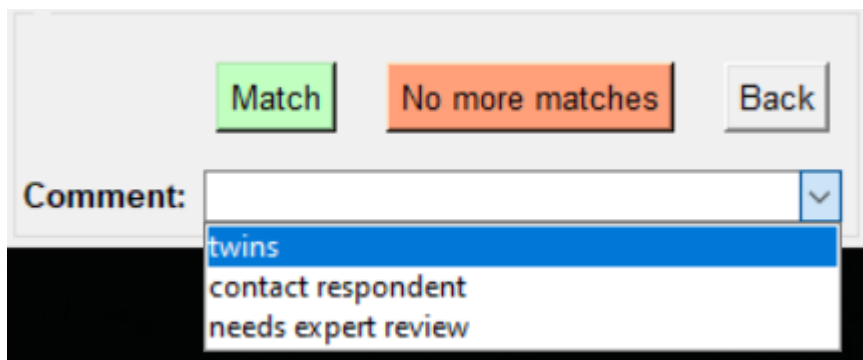
## [custom\_settings]

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.

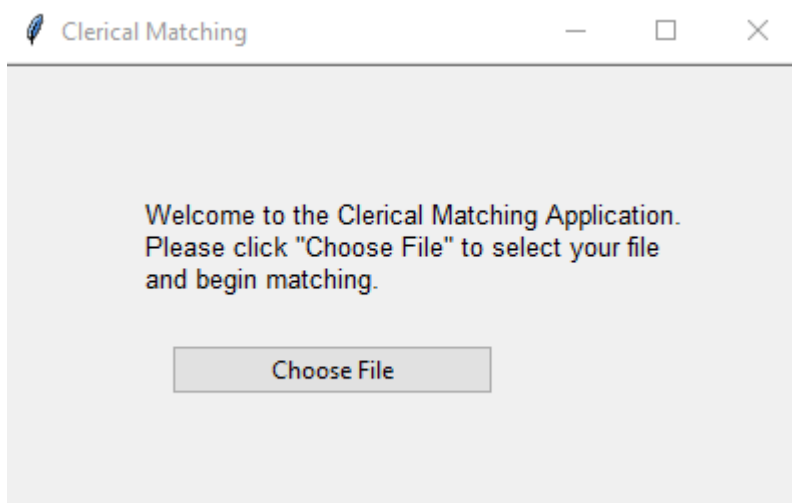
A screenshot of a web interface showing three buttons at the top: 'Match' (green), 'No more matches' (orange), and 'Back' (grey). Below the buttons is a label 'Comment:' followed by a text input field with a dropdown arrow on the right.

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

A screenshot of the same web interface as above, but the dropdown menu is open, showing three options: 'twins', 'contact respondent', and 'needs expert review'. The 'twins' option is highlighted in blue.

## 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.



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

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

Users can then use the checkboxes on the left-hand side of the window 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.

If users wish to add a custom comment, or a pre-determined comment about records in this cluster, a comment can be entered in the comment box (or selected from the drop-down tool) before a matching decision is made.

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
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
1	fbf58a86-62f7-477e-9ff5-79ca0226e534	PES	ELVIN	UWINEZA	44	b9aa7018-2416-423f-8ef1-05598be7f90c
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
9	a31f70af-7b8c-41b2-8aef-4d9c7265a6ed	PES	MICHELLE			300e8ab0-1560-4208-8693-01e4c73a80ee
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 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.

To get your outputs in a pairwise linked version (see below) run the CROW\_output\_updater.py script.

	A	B
1	puid_Census	puid_PES
2	b9aa7018-2416-423f-8ef1-05598be7f90c	fbf58a86-62f7-477e-9ff5-79ca0226e534
3	3926c771-fc2a-4b34-af20-436a76f3e7db	fbf58a86-62f7-477e-9ff5-79ca0226e534
4	d50a54b8-1ad8-456d-8365-d492020ccf98	40fe35cf-b29b-453d-8187-0659c60bf18a
5	d50a54b8-1ad8-456d-8365-d492020ccf98	1c70b116-7c89-47b9-912a-f1088ea97242
6	cee8b75a-a7ab-44bf-a5f0-954bd82ef667	47f3c7fd-c733-4fd0-b26d-2aed357a600c
7	9b46a536-8d25-4fd8-b658-e13b2630ed8c	1965026c-d580-4e6d-8002-674516572a52