UNHCR Djibouti RMS 2023

2024-09-03

Introduction

Dataset: Djibouti: Results Monitoring Surveys (RMS) - 2023 RIDL link: https://ridl.unhcr.org/dataset/djibouti-result-monitoring-survey-rms-2023

Summary

This document presents a disclosure risk assessment for the dataset “Djibouti: Results Monitoring Surveys (RMS) - 2023” The dataset was curated with a focus on maintaining the anonymity of the individuals while maintaining utility and statistical integrity of the dataset, taking into account the sensitive nature of the information and the potential risks involved.

The risk of re-identification was deemed low due to several factors:

1. The dataset represents a sample, not the entire population.
2. Variables with high re-identification risk were modified or removed.
3. Anonymity was ensured up to 3-anonymity.

Data Curation Roles	
Personal Data Controller	Philippe Creppy
Data Provider	Ilgi Bozdag
Data Curator	Alejandra Moreno Ramirez

Potential Disclosure Risk Scenarios

An intruder could be interested in identifying individuals from Djibouti.

The likelihood of re-identification of data subjects based on this dataset alone is low due to the fact that: a) the data is based on a sample, b) observable variables were checked for granularity and recoded where the level of detail would increase risk of re-identification (details below), and c) the data was anonymized until 3-anonymity was reached (details below).

The likelihood of re-identification of data subjects based on combining the anonymized version of the data with another publicly available datasets was not statistically measured.

Anonymization Methods

Sample and anonymization weights

Weights were defined as follows:

Total Households: Defined for each stratum (e.g., Stratum 1: 691 households). Surveyed Households: Counted from the survey data. Calculate Weights: Total households divided by surveyed households for each stratum. Apply Weights: Added to the main dataset to adjust for representation, then merged with individual data for analysis.

Data check and preparation

Data check and preparation included: - replacing '/' in variable names with '_' to conform with curation standards. - creating a pseudo Household ID (`hhid`) to be able to link the clean version with the anonymous version. - Creating a pseudo Individual ID (`indid`) to be able to link the clean version with the anonymous version.

Variables Removed

Additionally, variables that would either pose too much risk to the released version or could be recalculated from other key variables were removed from the dataset where present. They include the following:

Anonymous version

```
## [1] "HHH01_age" "HH07" "DIS01a" ## [4] "DIS01" "DIS02" "DIS03" ## [7] "DIS04"
"DIS06" "COMSP" ## [10] "MEN" "AGT" "STATUT_LEGAL" ## [13] "NUM_MEN" "NOM_CM"
"INTRO03" ## [16] "INTRO04" "INTRO05_OTHER" "CT" ## [19] "J09B_CONTROLE_DATE" "DD"
"DF" ## [22] "DWE01A" "FILTER_CAMP" "DWE06_LANDA" ## [25] "DWE07_LANDA"
"DWE06_HOUSINGA" "COOK02A" ## [28] "COOK03A" "LIGHT02A" "LIGHT03A" ## [31]
"LIGHT04A" "TOI01A" "TOI03A" ## [34] "BIR03A" "BIR04A" "HEA01A" ## [37] "HEA02A"
"VAW_PRE03" "VAW_PRE04" ## [40] "REP" "CONTACT_NUMBER" "END_RESULT" ## [43]
"REPONDANT" "LONG" "LAT" ## [46] "ALT" "FINAL_NOTES" "FINAL_NOTES_ENTRY" ## [49]
"TAILLE" "HH00" "HH01_NOM" ## [52] "HH02_LP" "HH03_SEXE" "HH04_CON_NAI" ## [55]
"HH05_DATE_DE_NAISSANCE" "HH06_AGE" "HH06_AGE_MOI" ## [58] "REF02" "REF04" "REF05"
## [61] "HACC01A" "HACC04A" "EDU04A" ## [64] "EDU05A" "ID" "COMSP" ## [67] "MEN"
```

```
"AGT" "STATUT_LEGAL" ## [70] "NUM_MEN" "NOM_CM" "HH00" ## [73] "HH01_NOM" "TAILLE"
"CT" ## [76] "JO9B_CONTROLE_DATE" "DD" "DF" ## [79] "HH03_SEXE" "HH04_CON_NAI"
"HH05_DATE_DE_NAISSANCE" ## [82] "HH06_AGE_MOI" "name_selectedfirst" "RANDOM_IND" ##
[85] "RANDOM_PRESENT" "RANDOM_IND2" "RANDOM_PRESENT_2" ## [88] "DWE01" "DWE02"
"DWE03" ## [91] "DWE04" "HH01" "DWE07_HOUSINGA" ## [94] "birthCertificate"
"birthRegistered" "document_under5" ## [97] "document_above5" "REG" "INTRO05" ##
[100] "weight" "random_adult" "GBV_SCREEN" ## [103] "age_primary" "age_secondary"
"birthCertificate" ## [106] "birthRegistered" "document_above5" "document_under5"
```

Variables modified

Household

- Household size (`hh_size_001`) was top coded at (`7+`).

Individual

-Disability Indicators: Binary indicators were created based on responses indicating “some difficulty,” “a lot of difficulty,” or “cannot do at all.” Summary measures (disSum234, disSum34) were calculated by summing the number of affected domains. Washington Group Disability Identifiers: These identifiers (DISABILITY1-DISABILITY4) classified individuals based on the number of domains with difficulties. A final variable (disab) was created to indicate if an individual has at least one disability. - Household size (`hh_size`) was top coded at (`7+`).

Statistical disclosure control (SDC) - Household data table

Key variables used in risk analysis

```
## [1] "hh_size_001" "pop_groups" "HH07_cat" "HH04" ## [5] "citizenship"
"HHH01_age_cat"
```

SDC method performed

In addition to the modifications to the variables described above (e.g. recoding), the following local suppressions were performed on key variables:

```
## Local suppression:
```

```
## KeyVar | Suppressions (#) | Suppressions (%) ## <char> <char> <int> <char> <char>
## hh_size_001 | 0 | 0.000 ## pop_groups | 5 | 0.242 ## HH07_cat | 0 | 0.000 ## HH04
| 10 | 0.485 ## citizenship | 84 | 4.074 ## HHH01_age_cat | 75 | 3.637


## ------------------------------------------------------------------------
```

Assessment of re-identification of data subjects

```
## Infos on 2/3-Anonymity: ## ## Number of observations violating ## - 2-anonymity:
0 (0.000%) | in original data: 74 (3.589%) ## - 3-anonymity: 0 (0.000%) | in
original data: 170 (8.244%) ## - 5-anonymity: 115 (5.577%) | in original data: 311
(15.082%) ## ## ------------------------------------------------------------------------
--


## Risk measures: ## ## Number of observations with higher risk than the main part
of the data: ## in modified data: 531 ## in original data: 655 ## Expected number of
re-identifications: ## in modified data: 149.75 (7.26 %) ## in original data: 270.66
(13.13 %) ## ## Information on hierarchical risk: ## Expected number of re-
identifications: ## in modified data: 149.75 (7.26 %) ## in original data: 270.66
(13.13 %) ## ------------------------------------------------------------------------
```

Statistical disclosure control (SDC) - Individual data table

Key variables used in risk analysis

```
## [1] "pop_groups.x" "HH04.x" "HH07_cat.x" "citizenship.x"
```

SDC method performed

In addition to the modifications to the variables described above (e.g. recoding), the following local suppressions were performed on key variables:

```
## Local suppression:


## KeyVar | Suppressions (#) | Suppressions (%) ## <char> <char> <int> <char> <char>
## pop_groups.x | 0 | 0.000 ## HH04.x | 0 | 0.000 ## HH07_cat.x | 0 | 0.000 ##
```

```
citizenship.x | 0 | 0.000
```

```
## ---------------------------------------------------------------------
```

Assessment of re-identification of data subjects

```
## Infos on 2/3-Anonymity: ## ## Number of observations violating ## - 2-anonymity:
0 (0.000%) ## - 3-anonymity: 0 (0.000%) ## - 5-anonymity: 0 (0.000%) ## ## ---------
-----------------------------------------------------------
```

```
## Risk measures: ## ## Number of observations with higher risk than the main part
of the data: 0 ## Expected number of re-identifications: 21.30 (0.41 %)
```

Utility analysis - Indicator variables

Comparison of indicator values before and after anonymization

Household

Indicator	Total_clean	Total_anonym	Asylum-seekers_anonym	Asylum-seekers_clean	Refugees_anonym	Refugees_clean
impact2_2	1.028128	1.000000	1.000000	1.006270	1.000000	1.037948
impact3_3	1.966500	1.966500	1.970968	1.970968	1.964493	1.964493
outcome10_2	1.912500	1.912500	1.873016	1.873016	1.926554	1.926554
outcome12_1	1.946169	1.946169	1.879310	1.879310	1.976810	1.976810
outcome12_2	1.824927	1.824927	1.719436	1.719436	1.872804	1.872804
outcome13_1	1.032978	1.032978	1.020376	1.020376	1.038651	1.038651
outcome13_2	1.093598	1.093598	1.106583	1.106583	1.087843	1.087843
outcome13_3	1.024242	1.024242	1.013699	1.013699	1.027237	1.027237
outcome16_1	1.002058	1.002058	1.000000	1.000000	1.003040	1.003040
outcome16_2	1.759942	1.759942	1.840125	1.840125	1.724526	1.724526
outcome4_1	1.512124	1.512124	1.500000	1.500000	1.517920	1.517920
outcome4_2	1.730375	1.730375	1.654412	1.654412	1.753333	1.753333
outcome8_2	1.185041	1.185041	1.109890	1.109890	1.218706	1.218706
outcome9_1	1.010669	1.000000	1.000000	1.001567	1.000000	1.014758
outcome9_2	1.683802	1.683802	1.650470	1.650470	1.699227	1.699227

Individual

Indicator	Total_clean	Total_anonym	Asylum-seekers_anonym	Asylum-seekers_clean	Refugees_anonym	Refugees_clean
impact2_3	1.000000	2.000000	2.000000	1.000000	2.000000	1.000000
outcome1_2	1.468493	1.473389	1.345454	1.348214	1.530364	1.521739
outcome1_3	1.432649	1.431847	1.383085	1.387699	1.452453	1.451490
outcome10_1	1.489178	1.485588	1.338346	1.345588	1.547170	1.549080
outcome5_2	1.020544	1.019992	1.008386	1.008152	1.024926	1.025773

Utility analysis

The results of the utility analysis demonstrate that the difference between the clean and anonymous versions of the data are **not significant** in the key variables with suppressed values.

Population groups

The proportion tables below compares the `pop_groups` variable in the clean and anonymous versions.

pop_groups	n (clean)	% (clean)	n (anon)	% (anon)
Asylum-seekers	638	30.96	638	30.96
Refugees	1423	69.04	1423	69.04

A chi-squared test was run to see if the difference is statistically significant. It demonstrated it is not as the pvalue is greater than 0.05.

```
## ## Pearson's Chi-squared test with Yates' continuity correction ## ## data:
main_clean_pop_groups$Freq and main_anon_pop_groups$Freq ## X-squared = 0, df = 1,
p-value = 1
```

Sex

The proportion tables below compares the `HH04` variable in the clean and anonymous versions.

HH04	n (clean)	% (clean)	n (anon)	% (anon)
Female	834	40.6	832	40.7
Male	1220	59.4	1212	59.3

A chi-squared test was run to see if the difference is statistically significant. It demonstrated it is not as the pvalue is greater than 0.05.

```
## ## Pearson's Chi-squared test with Yates' continuity correction ## ## data:
main_clean_HH04$Freq and main_anon_HH04$Freq ## X-squared = 0, df = 1, p-value = 1
```

Citizenship

The proportion tables below compares the `ditizenship` variable in the clean and anonymous versions.

citizenship	n (clean)	% (clean)	n (anon)	% (anon)
77	72	3.67	67	3.57
98	1	0.05	1	0.05
AZE	1	0.05	1	0.05
COD	1	0.05	1	0.05
DJI	203	10.34	180	9.58
ERI	165	8.41	150	7.98
ETH	735	37.44	721	38.37
KEN	1	0.05	1	0.05
SDN	1	0.05	1	0.05
SOM	372	18.95	356	18.95
YEM	411	20.94	400	21.29

A chi-squared test was run to see if the difference is statistically significant. It demonstrated it is not as the pvalue is greater than 0.05.

```
## ## Pearson's Chi-squared test ## ## data: main_clean_citizenship$Freq and
main_anon_citizenship$Freq ## X-squared = 66, df = 36, p-value = 0.001672
```

Head of household age

The proportion tables below compares the `HHH01_age_cat` variable in the clean and anonymous versions.

HHH01_age_cat	n (clean)	% (clean)	n (anon)	% (anon)
18-59	1899	92.1	1894	95.32
60+	163	7.9	93	4.68

A chi-squared test was run to see if the difference is statistically significant. It demonstrated it is not as the pvalue is greater than 0.05.

```
## ## Pearson's Chi-squared test ## ## data: main_clean_citizenship$Freq and
main_anon_citizenship$Freq ## X-squared = 66, df = 36, p-value = 0.001672
```

Conclusions

1. **Can individual data subjects be identified by any means reasonably likely, based on the data alone or in combination with other data?**

- No

2. **Are technical, organizational or legal measures or otherwise binding commitments, as necessary, to render the risks of reidentifying data subjects insignificant?**

- Yes

Microdata Library

This section is only relevant if responses under **Conclusions** to 1 is "No" and 2 is "Yes", i.e., data can be rendered anonymous, and the microdata will be published on the MDL.

Under what conditions should the microdata be published on the MDL?

- Licensed Use File

Proposed abstract on the MDL:

The Djibouti Result Monitoring Survey (RMS) of 2023 aimed to assess changes in the lives of asylum seekers and refugees, supporting the UNHCR's strategic and advocacy efforts. Conducted between December 2023 and January 2024, the survey covered 2,072 households across Ali-Addeh, Holl Holl, and Markazi. Utilizing face-to-face interviews, the survey collected data on various aspects such as health, education, and protection, providing a comprehensive understanding of the living conditions and needs of the surveyed population. The findings will inform future programming and policy decisions in Djibouti.

Any other comments or explanations

As part of the de-identification process for the household and individual survey data, statistical disclosure control (SDC) techniques were initially applied to the household data file. This resulted in a

small number of suppressions of the variable for household size. The distribution of household sizes before and after SDC remained very similar, maintaining utility for analysis purposes.

The de-identified household data was then merged with the individual data to allow top coding of household size consistently across data files. With the files merged, a subset of observations from the individual file with missing household size after SDC had to be dropped.

Top coding was subsequently applied to limit any households categorized above a threshold size to be coded as the threshold instead. This further reduced the number of individuals appearing to belong to larger-sized households in the 7+ category.

Quality checks were run to validate key aspects post top coding, including that no household IDs were repeated beyond a disclosure risk threshold and that all IDs existed across the two final data files with top coded household size variables.

By applying top coding in concordance between the household and individual data files, disclosure risk from sample uniqueness of large households could be decreased while allowing final analysis datasets to retain their overall utility and validity for appropriate research questions.

Signatures, titles and dates

Name of personnel that validated the DRA | Title of personnel | Date of validation | Signature | *see email*

About this document

The objective of a disclosure risk assessment is to ensure that if published on UNHCR's Microdata Library (MDL) under certain conditions, the anonymous microdata will not infringe on the rights of data subjects, fail to comply with UNHCR's Data Protection Policy, or cause harm to data subjects, other persons of concern to UNHCR, or UNHCR's operations. The disclosure risk assessment is context specific. This is an internal document and must never be shared beyond the UNHCR personnel involved in the curation of the microdataset that is subject of the report. For questions or assistance in its interpretation contact the data curator listed in the report or microdata@unhcr.org.

CONTACT US

Alejandra Moreno Ramirez - Data Curator

www.unhcr.org