Policy Research Working Paper 10631

Data Gaps in Microdata in the Context of Forced Displacement

Takaaki Masaki
B. Madson

WORLD BANK GROUP
Poverty and Equity Global Practice
December 2023

Policy Research Working Paper 10631

Abstract

This paper aims to understand the existing gaps in micro-level data on forcibly displaced people—refugees and internally displaced persons. The paper undertakes a comprehensive review of all existing micro-level data sets in the United Nations High Commissioner for Refugees Microdata Library and the World Bank Microdata Library. It first identifies a corpus of micro-level data sets that are designed to have a representative sample of refugees and/or internally displaced persons and assesses gaps in geographical and thematic coverage. The paper then evaluates whether the data sets contain a core set of questions that are essential for the proper identification of refugees and internally displaced persons. The findings show that microdata on forcibly displaced people are comparatively rich in Sub-Saharan Africa in contrast to other regions. However, data scarcity is notably pronounced in countries facing fragility and conflict. Scarcity is also evident among internally displaced persons and on topics such as labor and employment, finance (for instance, credit, debt, and banking), agriculture/livestock/fishery, and education. The paper also highlights that many of the existing micro-level data sets on forcibly displaced people do not contain the core set of questions needed for proper identification of refugees or internally displaced persons according to international statistical standards.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at tmasaki@worldbank.org.

Data Gaps in Microdata in the Context of Forced Displacement[*]

Takaaki Masaki[a]
B. Madson[b]

Keywords: Refugees, Internally Displaced Persons (IDPs), Forced Migration, Forced Displacement, Microdata
JEL: F22; C8

[*] The authors wish to thank Donatien Beguy, Patrick Brock, Craig Loschmann, Utz Johann Pape, Felix Schmieding, Nistha Sinha, Jeffery Tanner, and Domenico Tobasso for their comments and advice. We are also grateful to the participants at the workshop on "Filling the Remaining Data Gaps in the Forced Displacement Context" organized by the World Bank – UNHCR Joint Data Center on Forced Displacement (JDC). We appreciate technical support from Aivin Solatorio and Olivier Dupriez on the use of the nlp4dev tool. Finally, we express our gratitude to Benu Bidani, Aissatou Dicko, Björn Gillsäter, and Maja Lazic for their guidance and encouragement. This work is funded by the JDC and its partners: the Government of Denmark represented by the Danish Ministry of Foreign Affairs; the European Union represented by the Directorate-General for International Partnerships (INTPA) at the European Commission; and the United States Government represented by the U.S. Bureau of Population, Refugees, and Migration (PRM). The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, nor do they represent the views of the Executive Directors of the World Bank and the governments they represent, or the views of UNHCR.
[a] Senior Economist, World Bank / UNHCR Joint Data Center on Forced Displacement; the World Bank Poverty and Equity Global Practice, tmasaki@worldbank.org.
[b] Ph.D. candidate in Public Policy, Duke University, braydon.madson@duke.edu.

1. Introduction

The current era is marked by the highest number of forcibly displaced people (FDP) in history. As a result of the war in Ukraine coupled with ongoing conflict and disaster in countries such as Afghanistan, Ethiopia, Somalia, the Syrian Arab Republic, the República Bolivariana de Venezuela, and the Republic of Yemen, the number of people forced to flee due to fears of persecution, conflict, violence, human rights violations, and other events seriously disturbing public order had reached more than 110 million.[2] Understanding the extent and nature of the challenges they face is crucial for developing effective policy responses to address their needs and support their successful integration into their host communities or return to their places of origin or previous residence. However, existing data gaps on FDP make it challenging to design and implement such responses.

The aim of this paper lies in understanding existing gaps in microdata on FDP – refugees and internally displaced persons (IDPs). This paper does this by undertaking a review of all existing micro-level datasets in the UNHCR Microdata Library (MDL)[3] and World Bank (WB)[4] MDL from 53 low-income and middle-income countries with a significant presence of FDP.[5] The microdata in these databases are documented in compliance with international standards and practices. They contain a rich source of information on various attributes of datasets, including the country and dates of data collection, sampling strategy, survey modules, and other related aspects. While submission of datasets to the libraries is voluntary and thus does not guarantee an exhaustive list of all publicly available existing datasets on FDP, they are considered to be among the largest databases of microdata concerning development and forced displacement (Thompson 2010; EGRISS 2023).

For the purpose of this study, microdata are defined as primary data collected from household surveys. Our focus lies in identifying micro-level datasets that are publicly available and designed to have a representative sample[6] of FDP so that collected data can be disaggregated for refugees and/or IDPs specifically. By studying the geographical and thematic coverage of existing FDP microdata, we seek to also shed light on critical data gaps that remain to be filled with further data collection efforts.

The paper identifies critical gaps in geographical and thematic coverage as well as compliance with international recommendations for proper identification of displacement status. The paper highlights that microdata is comparatively rich in Sub-Saharan Africa in contrast to other regions. However, data scarcity is notably pronounced in countries facing fragility and conflict and also among IDPs. There are also certain topics that are relatively lacking in the FDP microdata. These include topics like labor, finance (e.g., credit, debt, banking), agriculture/livestock/fishery, and education whereas FDP datasets are relatively rich in health, food insecurity, and water and sanitation in addition to coping strategies and protection. Furthermore, a large majority of questionnaires for these existing micro-level datasets do not include a core set of questions needed for proper identification of displacement status. This calls for further efforts

[2] See https://www.unhcr.org/news/stories/unhcr-s-grandi-110-million-displaced-indictment-our-world.
[3] The UNHCR MDL is available at https://microdata.unhcr.org/index.php/about.
[4] The WB MDL is available at https://microdata.worldbank.org/index.php/about.
[5] We focus on low-income and middle-income countries because they account for roughly 95 percent of FDP (refugees + IDPs) in the world based on data from UNHCR Refugee Finder and IDMC.
[6] A representative sample means that when analyzed, the observed characteristics of the sample reflect the true characteristics in the target population that is being researched (Baal and Ronkainen, 2017).

to ensure that questionnaires are designed to comply with the international standards laid out by the Expert Group on Refugee, IDP and Statelessness Statistics (EGRISS).[7]

Our findings are relevant to the literature on knowledge gaps on FDP. Though few, there are prior studies undertaken to review existing evidence and data gaps around migrants and displaced populations (e.g., Baal 2021; Rico and Camilo 2022; USAID 2021; Berretta et al. 2023; EGRISS 2023). Berretta et al. (2023) provide a systematic review of existing studies on the causes of migration. Most recently, the *Compilers' Manual on Forced Displacement Statistics* by the Expert Group on Refugee, Internally Displaced Persons and Statelessness Statistics (EGRISS 2023) features some of the recent attempts to collect micro-level data on refugees and IDPs through censuses and household surveys.

This paper develops a systematic approach to identify publicly available micro-level datasets on FDP. While there are several previous studies that attempt to assess data gaps on FDP,[8] they are limited in scope in terms of both geographical and temporal coverage and often lack a rigorous and systematic approach to guard against reviewer selection bias (Sida 2020; James et al. 2016; Campbell Collaboration 2015; Clapton et al. 2015; Gough et al. 2012; Rico and Camilo 2022).[9] This current analysis systematically reviews all datasets on FDP from the UNHCR and WB MDLs. While our paper has its limitations, the proposed methodology is the first attempt to develop a replicable and scalable approach to identify the universe of publicly available FDP microdata.

Our approach is innovative in that we leverage the entire collection of datasets found in the UNHCR and WB MDLs with the aid of a natural language processing tool and text mining. This involves scraping metadata from thousands of publicly available datasets in those databases, evaluating whether they include a representative sample of refugees and IDPs, and assessing various attributes of each dataset (e.g., geographical coverage, thematic focus, kinds of questions available for identification of displacement status). The innovation in our approach lays the groundwork for future data gap mapping exercises to be more objective, replicable, and scalable.[10]

Additionally, this paper contributes to highlighting specific data gaps that have yet to be filled. To make further progress in filling data gaps, it is critical that we understand where those gaps exist today. This paper does exactly that by analyzing what and where data exists, so policy makers and development practitioners alike can gain a broad knowledge of the current data landscape on the forcibly displaced and aid their decision making based on available data.

2. Data

For this analysis, we leverage a rich catalogue of micro-level datasets from the UNHCR and World Bank MDLs. The WB MDL includes datasets from the World Bank, other international organizations, statistical agencies and other agencies in low- and middle-income countries. These datasets may also originate from population, housing or agricultural censuses or through an administrative data collection processes. This

[7] See more information on EGRISS at https://egrisstats.org/.
[8] See, for instance, Baal 2021; S 2020; Rico and Camilo 2022; USAID 2021; Berretta et al. 2023.
[9] For instance, Rico and Camilo (2022) assess the coverage of information on migrants by the censuses and regular household surveys, but their analysis is geographically confined to Latin America.
[10] All analysis for this report is performed in R. The replication codes are all available from https://github.com/takaakimasaki/DisplacementDataGap.

MDL contained 4,539 datasets at the time of this research (December 2022), some of which were also cross-listed in the UNHCR MDL.

The UNHCR MDL contains micro-level datasets that are of concern to UNHCR's mission and mandates. These datasets often – if not always – include a sample of refugees, asylum seekers, IDPs, or stateless people. The datasets come from censuses, registration and administrative exercises, and surveys. This MDL contained 546 datasets at the time of our mapping exercise.

One of the key advantages of using these databases is that they offer detailed metadata which help us understand the main characteristics of each dataset. The metadata contains a rich set of attributes pertaining to each micro-level dataset, including the country of data collection, the producer(s) of the dataset, brief description or abstract and thematic scope of a dataset, dates of data collection, unit of analysis (e.g., households, individuals), geographic coverage (e.g., national, regions, camps), data type (e.g., sample survey data, census, administrative), questionnaire modules, as well as sampling strategy. A complete list and explanation of each attribute in the metadata can be found in Annex B: Metadata.

In terms of geographical coverage, this study covers all low-income and middle-income countries with a significant presence of refugees, other Venezuelan refugees and migrants[11] in need of international protection, and/or IDPs. We apply 100,000 (in terms of the number of FDP as of 2021) as an inclusion threshold for this analysis. This results in the final list of 53 low-income and middle-income countries, which altogether account for 23 million refugees and 58 million IDPs.[12]

3. Methodology

The methodology we apply to analyze micro-level datasets in the UNHCR and WB MDLs follows a procedure that is systematic, replicable, and scalable (Figure 1). First, we scrape the metadata from all the micro-level datasets found on the UNHCR and WB MDLs.[13] We find 412 datasets from UNHCR MDL and 1,927 datasets from WB MDL that have been collected in the sample of countries under study.

Second, we remove *false positive* datasets from the UNHCR MDL. While a large majority of datasets in the UNHCR MDL are individual/household-level household survey datasets sampled from refugees or IDPs, some are not. Furthermore, many datasets do not have a clearly defined sampling frame from which a representative sample can be drawn, thus failing to meet our inclusion criteria for this exercise. A few datasets listed in the UNHCR MDL are indeed based on key informant interviews (instead of a

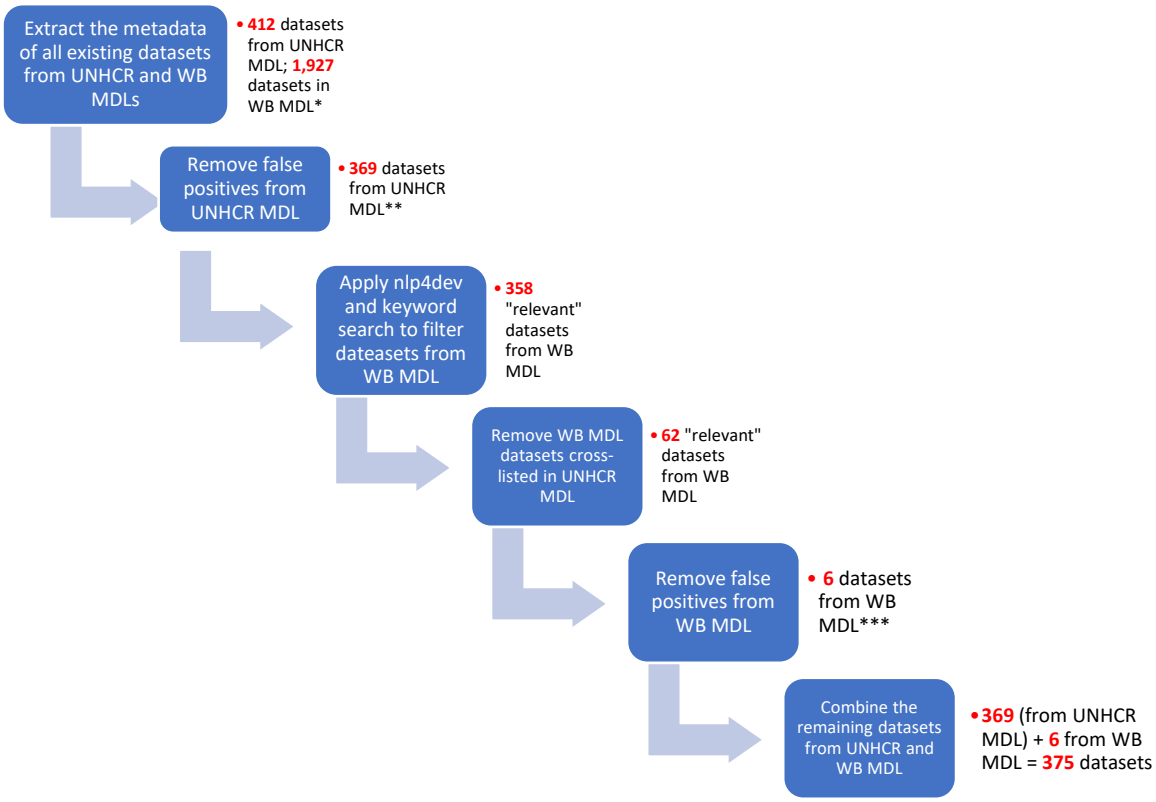
[11] Venezuelan refugees and migrants are included in this analysis due to their status as Persons of Concern (PoC) under UNHCR's mandate. The reason why migrants are included here is laid out by UNHCR: "Venezuelan [migrants] refers to persons of Venezuelan origin who are likely to be in need of international protection under the criteria contained in the Cartagena Declaration, but who have not applied for asylum in the country in which they are present. Regardless of status, Venezuelan [migrants] require protection against forced returns and access to basic services." (UNHCR 2020). In short, since Venezuelan migrants are likely to be in refugee-like situations, they are treated like refugees and thus should be included in this analysis.

[12] The number of refugees, Venezuelan refugees and migrants and IDPs (in 2021) is taken from World Bank *World Development Indicators* (WDI) (https://databank.worldbank.org/source/world-development-indicators), UNHCR Refugee Finder, and Global Internal Displacement Database by the International Displacement Monitoring Center (IDMC) (https://www.internal-displacement.org/database/displacement-data), respectively. See Annex A: List of countries included in this study.

[13] The extraction of metadata from the UNHCR and WB MDL was performed in December 2022. This involved scraping the JSON files of all existing datasets from both sources and compiling them into a master list.

representative sample of refugees or IDPs) or event/transaction datasets whose unit of analysis is not the household or individual. If datasets do not have a representative sample of refugees or IDPs or their unit of analysis is not the household or individual, these datasets are flagged as false positive and thus excluded from our study. We identify 43 false positive datasets from the UNHCR MDL and removing them leaves us with 369 datasets from the UNHCR MDL.

Figure 1: Methodology to identify FDP datasets from UNHCR and WB MDLs



* This refers to all micro-level datasets from the 53 low-income and middle-income countries analyzed for this study.
** Not all datasets in UNHCR MDL are micro-level datasets or sample explicitly from refugees or IDPs and should be manually removed.
**Most of the relevant datasets in WB MDL prove false positive because while having some references to relevant topics such as migration and displacement in the metadata (which is typically tagged as relevant in nlp4dev), they do not specifically mention sampling explicitly from refugees or IDPs.

Third, we identify micro-level datasets from the WB MDL that have a representative sample of refugees or IDPs. A large majority of datasets in the WB MDL do not sample specifically from refugees or IDPs, calling for a methodology to systematically filter these datasets and identify those that do have a representative sample of FDP. To this end, we apply three different filters to first identify “potentially relevant” datasets from the WB MDL:

a. Use a natural language processing tool called nlp4dev[14] to tag each dataset in the WB MDL by thematic areas and keep those ones that are flagged as relevant to the issues of migration, displacement and refugees;[15]
b. Keep those datasets that mention "refugee(s)", "internally displaced", "internal displacement", "IDP(s)", or "Venezuelan(s)" in either the title or sampling section of a dataset in the metadata;
c. Keep those datasets that have UNHCR as one of the data producers in the metadata.

If a given dataset passes at least one of these filters, it is flagged as "potentially relevant."[16] It is worth highlighting that these filters have been developed in an iterative process. We consulted with a number of country experts both from UNHCR and the World Bank to avoid *false negatives* (or those datasets that do in fact sample from refugees or IDPs but are not correctly identified).[17] Applying these filters identifies 358 potentially relevant datasets from the WB MDL.

Fourth, we perform a manual verification to remove those datasets that are also cross-listed in the UNHCR MDL. It is worth noting that most of the potentially relevant datasets taken from the WB MDL are also already cross-listed in the UNHCR MDL and thus are removed to avoid double-counting. After removing these cross-listed datasets, we have 62 datasets left from the WB MDL. Fifth, we apply the same manual verification to identify false positive datasets from the WB MDL following the same procedure as done for the datasets from the UNHCR MDL, which ultimately leaves us with 6 datasets from the WB MDL.[18] Finally, we combine the identified 369 datasets from the UNHCR MDL and 6 datasets from the WB MDL to arrive at our final list of 375 FDP micro-level datasets.

After identifying the corpus of FDP datasets – or micro-level datasets that have a representative sample of refugees and IDPs – we then also manually code various attributes of each dataset that are not readily available from the UNHCR and WB MDLs. First, we evaluate whether a given dataset is *project-specific*. More specifically, we look at whether the sampling frame in a given dataset is narrowly defined to only include beneficiaries of a certain project or program. While project-specific datasets are useful in their

[14] The NLP4Dev application – developed by the WB Development Data Group – is a tool for applying Natural Language Processing (NLP) topic and word embedding models to improve document and data discoverability. These models are used to implement a semantic search engine and to generate summary, structured information on the thematic and geographic composition of a large corpus of unstructured documents. See more details in https://www.nlp4dev.org/.

[15] nlp4dev allows users to tag texts by 75 different topic areas and reports what share of the texts cover which topic areas. We use topic 39 – which includes a cluster of issues related to the issues of our concern including refugee, program, country, migration, migrant, labor, remittance, population, international, asylum – to tag potentially relevant datasets. More specifically, we extract the entire metadata from each dataset, turn it into a text file, pass it through nlp4dev to report the coverage rate of each topic and then if the coverage rate is above 10 percent, flag the dataset as potentially relevant.

[16] Of the datasets from the WB MDL, 11 percent have titles or sampling sections in languages other than English within the metadata. In these instances, we utilize the DeepL API (using the deeplr package in R) to translate them into English. Subsequently, we apply the same filters. Furthermore, 5.5 percent of datasets (*N*=107) have missing sampling sections and for these instances, we manually verify whether refugees or IDPs are included as a sample based on desk research.

[17] For example, just applying nlp4dev as a filter turned out to be insufficient for correctly identifying many of the IDP datasets and additional filters were added after consulting with country experts from UNHCR and the World Bank Poverty & Equity Global Practices.

[18] Most of the datasets flagged as relevant in the nlp4dev tool turn out to be false positive. They are mostly flagged as relevant because they either have a module specifically on migration or mentioned some key terms related to migration but turn out not to sample specifically from refugees or IDPs.

own right for evaluating or monitoring the impact or outcomes of a certain intervention, the data collected through such a survey cannot be used to generate reliable statistics for any population outside the beneficiaries (or control group if any) of that particular project or program. Examples of project-specific datasets include energy monitoring framework survey, program monitoring beneficiary survey, post-distribution monitoring survey, among others. As we note in the following section, a non-negligible share of the existing FDP datasets turns out to be project specific.

Second, we code the topic coverage of each dataset based on its questionnaire modules. The metadata of each dataset contains information on survey modules or main topics covered in the questionnaires (e.g., health, education, food insecurity). Since there is no pre-determined list of thematic areas commonly included in the corpus of FDP micro-level datasets that are known to us *ex ante*, we need to create our own taxonomy of topic areas. To this end, we look at keywords that are frequently mentioned in the descriptions of the survey modules in the metadata, rank-order them based on frequency, and then map them into 14 different broad topic areas. For instance, words such as "consumption", "income", "expenditure", "non-food", and "welfare" are clustered under the topic area of *consumption and welfare* and if a given dataset mentions at least one of these keywords in the description of the survey modules in its metadata, the dataset is coded as relevant to that topic area.[19] To assess gaps in topic coverage among FDP datasets, we also apply the same methodology to all other micro-level datasets in the WB MDL and compare differences in topic coverage between FDP datasets as identified in our analysis and other non-FDP datasets contained in the WB MDL.

Third, we also evaluate whether a given dataset contains the core set of questions recommended by EGRISS.[20] EGRISS recommends a set of questions that are essential for proper identification of refugees and IDPs through the International Recommendations on Refugee Statistics (IRRS)[21] and the International Recommendations on Internally Displaced Persons Statistic (IRIS),[22] respectively. For refugees, the identification questions recommended by EGRISS include (EGRISS 2018):

a) Country of birth
b) Country of citizenship
c) Acquisition of citizenship
d) Year or period of arrival in the country
e) Reason for migration

For IDPs, these are (EGRISS 2020):

a) Place of birth
b) Date of first displacement
c) Date of most recent displacement
d) Main reason for initial displacement
e) Main reason for most recent displacement
f) Place of usual residence
g) Place of habitual residence

We download questionnaires for each dataset from the UNHCR/WB MDLs and review each question in the questionnaires to assess whether they include any of the core questions listed above.

---

[19] See Annex C: Taxonomy of for details on the procedure taken to create a list of topic areas examined in this study.
[20] EGRISS is a multi-stakeholder group that was established by the United Nations Statistical Commission (UNSC) in 2016 and now consists of members from 56 national authorities and 36 regional and international organizations.
[21] The full document of IRRS is available at https://egrisstats.org/recommendations/international-recommendations-on-refugee-statistics-irrs/.
[22] The full document of IRIS is available at https://egrisstats.org/recommendations/international-recommendations-on-idp-statistics-iris/.

Finally, we also apply text mining to distinguish those datasets sampling from IDPs from those that only sample from refugees or Venezuelan refugees and migrants. We evaluate whether some keywords indicating the inclusion of IDPs in the sampling frame are mentioned in the sampling section of the metadata. These keywords included “internally displaced”, “internal displacement”, “IDP”, or “IDPs.” We then perform a manual verification to make sure that IDPs are indeed included in the sampling frame to generate representative data for this population.

4. Results

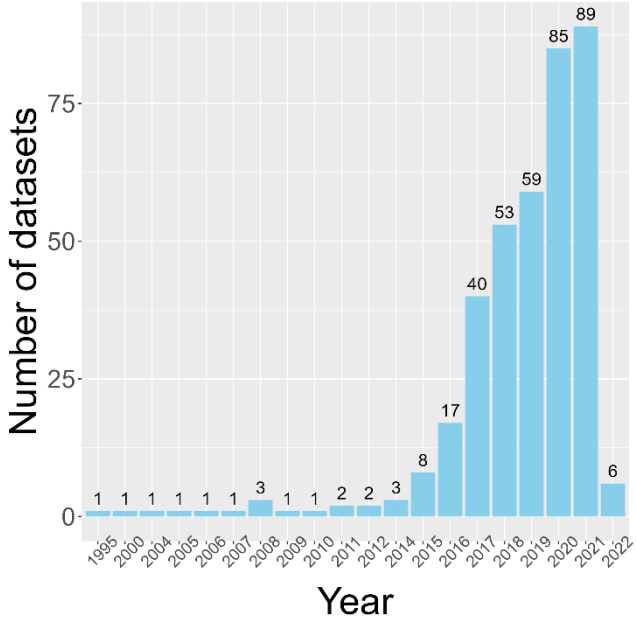
Through the methodology described above, we find 375 publicly available micro-level datasets that have a representative sample of refugees and/or IDPs.[23] Figure 2 plots the number of datasets over time between 1995 and 2022.[24] There has been a rapid increase in the number of FDP micro-level datasets collected and published in the UNHCR/WB MDLs especially since 2015. It is worth noting that part of this upward trend may also be explained by the fact that older micro-level datasets are not stored or published in the UNHCR/WB MDLs. For instance, the UNHCR MDL was launched in 2019 and despite efforts to retrospectively cover and publish older micro-level datasets collected before the inception of the UNHCR MDL, the extent to which older datasets are still missing from the database is unknown.[25]

[23] The results and codes for the analysis are all available from https://github.com/takaakimasaki/DisplacementDataGap.

[24] The oldest dataset available in WB/UNHCR MDL was collected in 1995 at the time of this study.

[25] See more recent developments in UNHCR MDL: https://www.unhcr.org/blogs/responsible-and-timely-sharing-data-on-unhcrs-microdata-library-in-2022/.

Figure 2: The number of FDP micro-level datasets over time



Notes: this graph shows a trend in the number of FDP micro-level datasets in the UNHCR/WB MDLs between 1995 and 2022 based on the year of data collection. Note that at the time of this study, many micro-level datasets collected in 2022 were not yet incorporated into the UNHCR/WB MDLs and thus the number for that year should be treated with caution.

Project specific datasets

Of the 375 FDP datasets, a non-negligible share of them turns out to be project specific, meaning that despite their utility for evaluating and monitoring the impact of a certain project or program, its application for analyzing any broader population of interest is quite limited. We find that 39 percent of the identified datasets sample from a narrow base of beneficiaries from a certain program or project.

It is important to draw a representative sample of a broader population of refugees and/or IDPs instead of sampling only from a narrow base. If the sample consists of only beneficiaries from a certain project or program, the generalizability of findings or conclusions drawn from that sample is very much limited and does not extend beyond the small confine of the target population that is relevant only to the project or program itself but not to broader communities of policy makers, development practitioners and scholars alike.

Geographical gaps

We identify a key set of country-level characteristics that may be correlated with the number of publicly available FDP datasets. These characteristics include GDP per capita (log-transformed) (*GDP pc (ln)*), the number of FDP (log-transformed) (*FDP (ln)*), as well as battle-related deaths (log-transformed) as a measure of conflict and fragility (*Battle-related deaths (ln)*).[26] We also include regional dummies to capture those regions that are underrepresented after accounting for those baseline country characteristics. We apply a negative binomial regression for this analysis as our dependent variable is the count of FDP datasets in a given country.

[26] Data on GDP pc and battle-related deaths are from WDI and FDP from WDI, UNHCR and IDCM. GDP pc is averaged for the period of 2010-2021 and the number of reported battle-related deaths is summed for the same period.

While countries hosting a greater number of FDP tend to have a greater number of publicly available FDP datasets, some countries defy this overall pattern – having no FDP dataset despite a relatively large size of FDP they host. Figure 3 Panel A plots the relationship between the size of FDP (or the total of refugees, Venezuelan refugees and migrants, and IDPs combined) and the number of publicly available FDP datasets.[27] The upward slope seen in the plot indicates that overall there is a positive correlation between the number of FDP and the number of publicly available FDP datasets. Bangladesh, Cameroon, Kenya, Lebanon and Uganda top the list in terms of the number of publicly available FDP datasets. It is worth highlighting that there are a number of countries that, despite having a sizable number of refugees or IDPs, have no FDP dataset identified through this analysis. Notable cases include the Syrian Arab Republic, Türkiye, Yemen, Pakistan, and China where the number of FDP exceeds 1 million, but no publicly available FDP dataset is found in the MDLs.

FDP data is also relatively lacking among countries in fragile and conflict-affected situations (FCS). Indeed, the level of conflict intensity is negatively correlated with the number of publicly available FDP datasets, meaning that the more fragile and violent a country, the fewer datasets it has available in the MDLs. This pattern is certainly not unique to FDP microdata. For instance, comparable poverty measures over time are often lacking in FCS countries (World Bank 2021).

A confluence of factors explains data deprivation in FCS countries. To inform discussions about policy responses towards FDP in a given country, the surveys must be representative at least for the target population of interest to policy makers. However, a reliable sampling frame is often missing in such conflict situations where population census rarely takes place, and registration data may also be obsolete, thereby complicating the implementation of sample household survey (Aguilera et al. 2020). Furthermore, collecting data in the fragile context also raises a concern for the safety of enumerators, inhibiting survey implementation in the unsafe regions (Corral et al. 2020). Furthermore, FCS countries also suffer from limited statistical capacity and a lack of resources to implement data collection (Cas et al. 2022).

In terms of the regional representation of publicly available FDP datasets, data is relatively rich in Sub-Saharan Africa in contrast to other regions.[28] In fact, Sub-Saharan Africa accounts for 53 percent of all FDP micro-level datasets (excluding project-specific datasets and those datasets collected before 2010). These patterns hold even after accounting for the above-mentioned baseline country characteristics such as GDP per capita, conflict intensity, and number of FDP (see Figure 3 Panel B). The average marginal effects of regional dummies are all negative and significant (except for South Asia) using Sub-Saharan Africa as a benchmark.

Data is particularly rich in East Africa. This skew is primarily driven by three countries that have a higher number of datasets than expected given the number of FDPs they host: Uganda, Kenya, and Tanzania. This pattern is reflective of a broader data and evidence landscape in Africa. Indeed, these three countries are also among the top in the region with the largest number of general micro-level datasets in the WB MDL. Additionally, Kenya and Uganda are among the most extensively studied countries in impact

[27] For this country-level analysis, we exclude those micro-level datasets that are collected prior to 2010 or project-specific datasets. Furthermore, some survey datasets are stored as separate entries in the WB and UNHCR MDL even though they are indeed part of the same survey (e.g., entries by camp or by wave). Those independent entries are collapsed as one when they are part of the same survey.

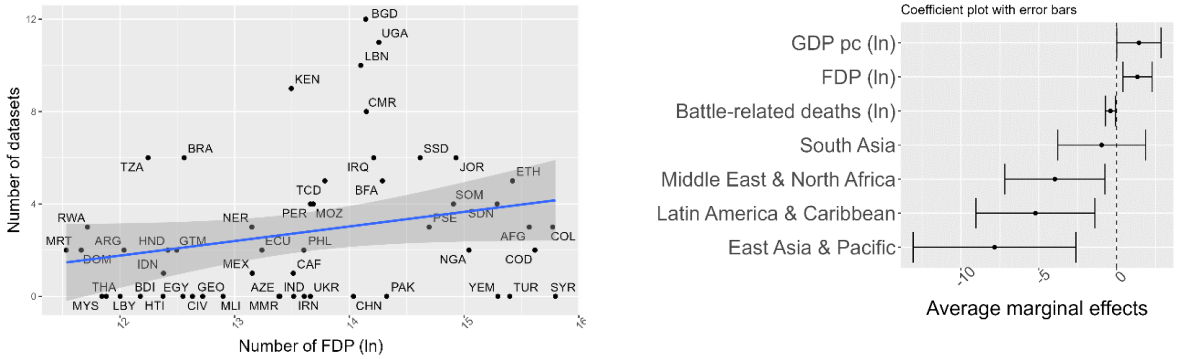
[28] Throughout this paper, we adopt the regional classification of the World Bank: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

evaluations (Sabet and Brown 2018). This trend in the broader academic literature helps develop the infrastructure within countries that is useful in facilitating the emergence of other research and datasets, like FDP datasets. In contrast, data are relatively scarce in Western/Central Africa, including Côte d'Ivoire, Central African Republic, Democratic Republic of Congo, Mali, and Nigeria as seen in Figure 3 Panel A.

Figure 3: Number of FDP datasets by country characteristics

A) Number of datasets and number of FDP

B) Country characteristics correlated with number of datasets



Notes: Panel A shows the bivariate relationship between the number of FDP datasets and number of FDP (log-transformed). ISO 3-letter country codes are also shown in the scatterplot. Panel B shows the estimated average marginal effects of each variable based on the negative binomial regression. In the regression model, regional dummies are included where sub-Saharan Africa is a dropped category. Note that some survey datasets are stored as separate entries in the WB and UNHCR MDL even though they are indeed part of the same survey (e.g., entries by camp or by wave). Those independent entries are collapsed as one when they are part of the same survey. Note that Europe and Central Asia is dropped from the negative binomial regression analysis because there is no dataset available from that region and thus no variability in the outcome variable. Lastly, this analysis only considers those FDP datasets that are non project specific and are collected after 2010 (*N*=221)[29] because any dataset that is project specific and/or collected before 2010 likely would have limited use for policy makers and development practitioners.

Publicly available micro-level datasets on IDPs

Overall, publicly available micro-level datasets for IDPs are largely lacking. There are only 22 datasets[30] in the UNHCR and WB MDLs that have a representative sample of IDPs. This is striking given that IDPs today account for about 60 percent of 110 million FDP across the world.[31] The Syrian Arab Republic, Democratic Republic of Congo, Colombia, Yemen, Ethiopia, and Türkiye all have more than 1 million IDPs and yet have no publicly available FDP datasets designed to sample specifically from IDPs in the MDLs (Figure 4).

There are a number of factors that make it particularly challenging to obtain representative data on IDPs. First, in most IDP contexts, no registration systems or other list of the total IDP population exists that can serve as a reliable sampling frame for a sample household survey. Where they do exist, they are often incomplete due to the high cost of updating them, as well as various other reasons (Baal and Ronkainen 2017). The collection of data on IDPs is further complicated by the fact that most IDPs are found in some

[29] The full catalogue of 221 micro-level datasets used for this analysis is available at https://rstudio.unhcr.org/FDMAtlas/.

[30] Note that some survey datasets are stored as separate entries in the UNHCR and WB MDLs even though they are indeed part of the same survey (e.g., entries by camp or by wave). Those independent entries are counted as one dataset when they are part of the same survey, thus resulting in a total number of 22 IDP datasets excluding those datasets that are project specific or collected before 2010.

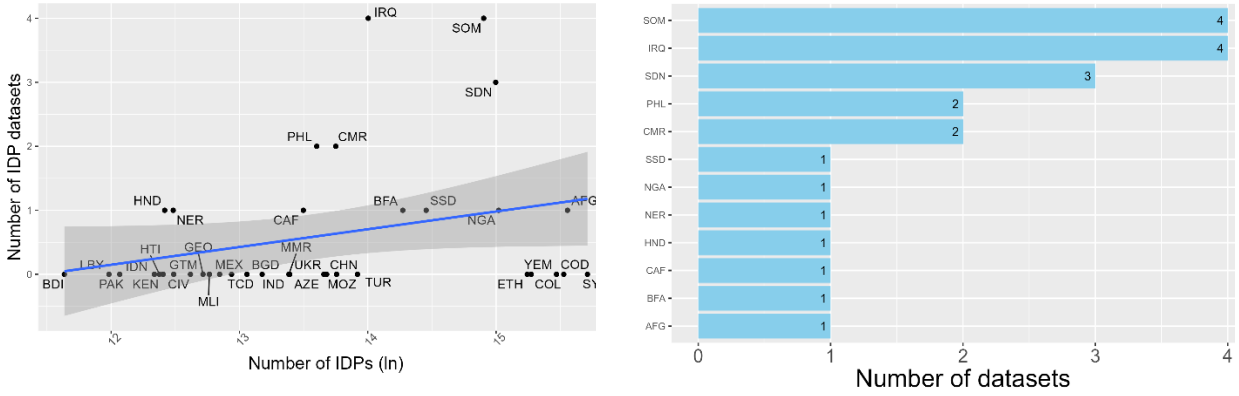
[31] See https://www.unhcr.org/global-trends.

of the most fragile countries where due to security concerns, physical access to reach them is restricted. Furthermore, many IDPs do not live in camps but rather mix with other population groups particularly in urban areas, thereby making it even more difficult to identify them (Baal and Ronkainen 2017).

Figure 4: Number of IDP datasets vis-à-vis size of IDPs hosted by country

A) Number of IDP datasets and number of IDP (ln)

B) Number of IDP datasets by country



Notes: Panel A shows the bivariate relationship between the number of IDP datasets and number of IDPs (log-transformed). Panel B shows the ranking of countries in terms of the number of IDP datasets available in the MDLs. Note that some survey datasets are stored as separate entries in the UNHCR and WB MDLs even though they are indeed part of the same survey (e.g., entries by camp or by wave). Those independent entries are counted as one dataset when they are part of the same survey, thus resulting in a total number of 22 IDP datasets excluding those datasets that are project specific or collected before 2010.

Topic coverage

In terms of the coverage of topics, FDP datasets are relatively richer in topics such as coping mechanisms, protection, water and sanitation, food insecurity, and health. Figure 5 Panel A illustrates the percent of FDP datasets that cover various topics in their questionnaire modules and Panel B shows differences in the percent of datasets covering each topic between FDP datasets and all other non-FDP datasets stored in the WB MDL.[32] The topic of health is by far the most common with roughly 78 percent of datasets mentioning keywords related to health in their descriptions of the questionnaire modules. Other common topics include food insecurity (49 percent) and water sanitation (39 percent). Less frequently included are topic areas related to labor and employment (19 percent); security and conflict (18 percent); finance, credit and debt (14 percent); agriculture and livestock (10 percent), and shocks (9 percent). The fact that topics such as security, conflict and shocks are often not included in the survey modules is not unique to FDP datasets. In fact, those topics are also relatively uncommon in other micro-level household surveys available in the WB MDL.

When compared to the topic coverage of all non-FDP micro-level datasets in the WB MDL, FDP datasets are scarcer in such topics as labor, finance, agriculture/livestock/fishery, and education. As shown in Figure 5 Panel B, these topics are much more likely to be picked up in the non-FDP micro-level datasets. When it comes to labor and employment in particular, the gap is over 50 percentage points illuminating

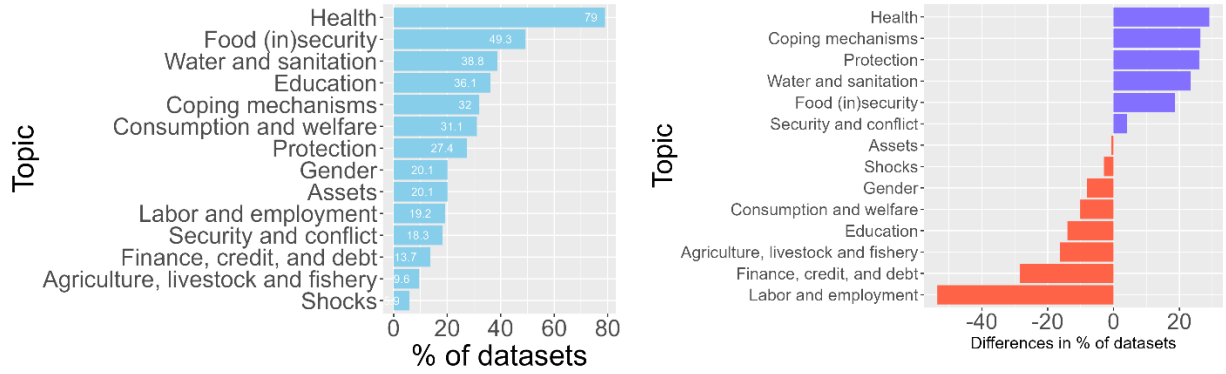
[32] See Annex C for further details on the methodology used to analyze topic coverage. For this topic coverage analysis, we exclude those datasets that are project specific or collected before 2010.

a significant relative data gap in FDP microdata on this topic area.[33] The World Bank's 2023 *World Development Report* highlights how such data will be critical for better understanding labor dynamics among FDP, including their skills and attributes, participation in the labor market, effects on productivity, among others (p. 61). A dearth of data on this topic limits our ability to unpack both the process and impact of labor market integration for FDP – a topic that has gained a lot of traction in both the policy and academic community (Ginn 2023).

Figure 5: Prevalence of Topics in Datasets

A) Percent of FDP datasets covering a particular topic in the survey modules

B) Differences in percent of datasets covering a particular topic in the survey modules between FDP datasets and non-FDP datasets in the WB MDLs



Notes: Panel A shows the percent of non-project-specific FDP datasets collected after 2010 covering each of the 14 different topic areas. See Annex C: Taxonomy of Topic Areas for further details on the procedure used to code the topic area of a dataset. Note that these policy areas are not mutually exclusive and a dataset can cover multiple different topic areas. Panel B shows the differences in the percent of datasets covering these topic areas between FDP datasets and all non-FDP datasets collected after 2010 in the WB MDL (*N*=726).

EGRISS International Standards

IRRS and IRIS provide a set of specific recommendations that countries and international organizations can use to improve the collection, collation, disaggregation, reporting, and overall quality of statistics on FDP. These recommendations are also intended to help improve national statistics on the stocks and flows and characteristics of FDPs, and to help make such statistics comparable internationally.

Core questions for proper identification of refugees and/or IDPs as recommended by EGRISS are often missing from FDP datasets. Figure 6 Panel A shows the percent of FDP datasets containing each of the core identification questions recommended for refugees[34] whereas Panel B shows the IDP equivalent of that. As seen in Panel A, the question of country of birth or citizenship is often missing from a large majority of datasets. Less than 40 percent of the datasets ask questions on citizenship and less than 30 percent of the datasets ask a question about the country of birth. While one is often substituted for the

[33] We find that there are a number of datasets whose survey modules contain keywords referring to child labor (e.g., "child labor" and "travail des enfants") flagged as relevant to labor and employment in non-FDP datasets in the WB MDL. However, these cases should be distinguished from those that have more standard questions around labor and employment and thus are not included when we aggregate the number of datasets for labor and employment.

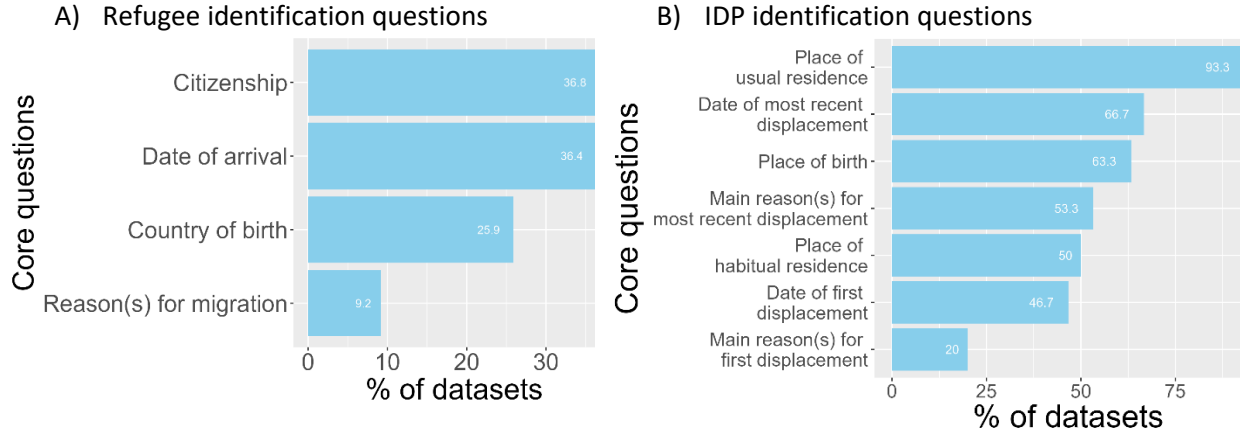
[34] For this analysis, we only included those FDP datasets designed to sample from refugees that are not project specific and collected after 2010.

other in many of these datasets,[35] the country of birth and citizenship are not the same and IRRS recommends that a survey captures both the country of birth and nationality (or nationalities) for both the legal and statistical identification of refugees or refugee-like persons.[36]

Reasons for displacement are also not commonly asked in FDP datasets. Of those datasets that are specifically designed to sample from refugees, reasons for (cross-border) migration are only asked in less than 10 percent of the refugee datasets. As for IDP datasets, only half of them ask reasons for the most recent displacement whereas questions around reasons for initial displacement are largely missing (with only 20 percent asking such questions). For refugees, reasons for migration should be included to properly identify those who have crossed an internationally recognized border on international protection grounds. For IDPs, reasons for initial and most recent displacements should be asked to identify those who have fled their place of habitual residence due to displacement-related protection needs and vulnerabilities such as "the effects of armed conflict, situations of generalized violence, violations of human rights, or natural or human-made disasters" (EGRISS 2023, p. 13).

The adaptation of future surveys to EGRISS recommendations is yet to be determined. Considering that these statistical standards and recommendations were established in recent years, it may not be surprising that the overall level of compliance with EGRISS recommendations is still evolving.

Figure 6: Prevalence of Identifying Questions in FDP Datasets



Notes: Panel A shows the percent of FDP datasets containing each of the core identification questions recommended by EGRISS for refugees (in all publicly available datasets designed to sample from refugees that are not project specific and collected after 2010) whereas Panel B shows IDP equivalent of that. For refugees, we do not make a distinction between the question on the country of citizenship and the acquisition of citizenship because many datasets we study often collapse these questions into one by simply asking respondents to indicate their nationality (or nationalities).

5. Conclusion

With an ever-growing number of people forced to flee their homes due to conflict, violence, fear of persecution, and human rights violations, data on FDP has become ever more important. A solid evidence base is crucial to inform effective policy responses to address their challenges and to establish such a base,

[35] More than 95 percent of the FDP datasets asked questions on either the country of birth or citizenship.

[36] While the country of birth does not change, nationality or citizenship can. And it is the country of citizenship or acquisition of citizenship that determines whether refugees and refugee related persons should still be regarded as refugees by the national authorities even though their country of birth may be different from that of their citizenship or nationality and how citizenship interplays with refugee status varies by country (EGRISS 2019, p. 28).

data – particularly representative micro-level data – on FDP will be essential. Despite the growing need for such data, however, the supply of it may not be adequately catching up.

This study leveraged a rich catalogue of micro-level datasets from the UNHCR and WB MDLs to undertake a stocktaking of all available micro-level datasets on FDP, thereby also shedding light on existing data gaps. By geography, the study finds that FDP datasets are relatively rich in Sub-Saharan Africa compared to other regions given that more than half of the publicly available FDP datasets identified through this study come from Sub-Saharan Africa. By topic, FDP datasets are relatively scarcer in topics such as labor and employment, finance, agriculture/livestock/fishery, and education.

One of the surprising results revealed from this study is an overall lack of publicly available microdata on IDPs. Of the 375 FDP datasets identified through this exercise, only 31 of them have a representative sample of IDPs. Furthermore, some countries hosting at least 1 million IDPs have zero publicly available dataset from the UNHCR and WB MDLs, including Colombia, Democratic Republic of Congo, Ethiopia, the Syrian Arab Republic, Türkiye, and Yemen.

The study also reveals how a large majority of the existing datasets did not contain a core set of questions for proper identification of refugees and IDPs. Some of the essential identification questions are often missing, such as the place or country of birth as well as reasons for and history of displacement. The omission of these core questions makes it difficult to properly identify the displacement status of sampled households or individuals. Since the international recommendations by EGRISS were established in recent years, it still remains to be seen to what extent future surveys will adapt to these recommendations.

These findings come with some caveats. First, the metadata of the datasets was extracted in December 2022. This means that our data gap map may already be missing datasets that have been added since then, calling for periodic updates in the analysis of data gaps. Second, our datasets only capture publicly available microdata from the UNHCR and World Bank MDLs. While we believe these MDLs contain the most datasets pertinent to microdata on FDPs, there may be other datasets not contained in either of these databases. Finally, as with all automated processes, it is possible that some datasets are miscoded due to the fact that the underlying metadata we use for this study is sometimes incomplete. We attempt to mitigate this concern by manual verification and making corrections in the metadata itself but there still remains more work to be done to ensure that the metadata is thoroughly populated in the UNHCR and WB MDLs.

Despite the limitations, this paper contributes to advancing our knowledge about existing data gaps on FDP and by developing a systematic process to identify them based on the information from the UNHCR and WB MDLs. Future data collection efforts may build upon the findings from this report and strategically target those geographical or thematic areas where data gaps still abound. Furthermore, the methodology developed in this study can be adopted, modified, and applied in other contexts, thereby making it easier to undertake this sort of a data gap mapping exercise.

Annex A: List of countries included in this study

The table below shows the list of all countries included in this study.

Table A 1: List of countries included in this study.

Countries with a significant presence of refugees (>100,000)	Countries with a significant presence of Venezuelan migrants (>100,000)	Countries with a significant presence of IDPs (>100,000)
Bangladesh	Argentina	Afghanistan
Cameroon	Brazil	Azerbaijan
Chad	Colombia	Bangladesh
Congo, Dem. Rep.	Dominican Republic	Burkina Faso
Egypt, Arab Rep.	Ecuador	Burundi
Ethiopia	Peru	Cameroon
India		Central African Republic
Iran, Islamic Rep.		Chad
Iraq		Colombia
Jordan		Congo, Dem. Rep.
Kenya		Ethiopia
Lebanon		Georgia
Malaysia		Guatemala
Mauritania		Haiti
Niger		Honduras
Pakistan		India
Palestine		Indonesia
China		Iraq
Rwanda		Côte d'Ivoire
South Sudan		Kenya
Sudan		Libya
Syrian Arab Republic		Mali
Tanzania		Mexico
Thailand		Mozambique
Türkiye		Myanmar
Uganda		Niger
		Nigeria
		Pakistan
		China
		Philippines
		Somalia
		South Sudan
		Sudan
		Syrian Arab Republic
		Türkiye
		Ukraine
		Yemen, Rep.

Annex B: Metadata

Table B 1 shows each variable coded in the metadata of each FDP dataset identified in UNHCR/WB MDL.

Table B 1: Metadata

Name	Description	Source
id	Unique numeric ID	UNHCR/WB
idno	Unique string ID	UNHCR/WB
nation_abbreviation	ISO-3 country abbreviation	UNHCR/WB
statement_title	Dataset title	UNHCR/WB
data_url	URL	UNHCR/WB
nation_name	Country name	UNHCR/WB
year	Year of data collection	UNHCR/WB
producers_name	Name(s) of data producer(s)	UNHCR/WB
entity_name	Name(s) of entity/entities involved in data collection	UNHCR/WB
entity_contact	Contact information of involved entities	UNHCR/WB
abstract	Brief description and scope of the dataset	UNHCR/WB
coll_dates	Year, month, and day data collection began and ended	UNHCR/WB
analysis_unit	Unit of analysis or data	UNHCR/WB
geog_coverage	Geographical coverage	UNHCR/WB
data_kind	Type of survey (e.g., sample survey, census, event, multi-frame, no sampling)	UNHCR/WB
notes	Description of topics and modules covered in the survey	UNHCR/WB
method_data_collectors_name	Method of data collection	UNHCR/WB
method_sampling	Description of sampling	UNHCR/WB
method_coll_mode	Mode of data collection (e.g., face-to-face, CAPI, CATI, Telephone, other)	UNHCR/WB
method_weight	Description of sampling weights	UNHCR/WB
topic_id	Numerical identifiers for which topics are covered in the survey based on nlp4dev (https://www.nlp4dev.org/)	Authors
topic_words	Text identifiers for which topics are covered in the survey based on nlp4dev (https://www.nlp4dev.org/)	Authors
topic_main	Indicator of the main topic covered in the survey based on nlp4dev (https://www.nlp4dev.org/)	Authors
project_specific	Binary indicator of whether the dataset is project specific or not: Yes or no	Authors
idp	Binary indicator of whether the dataset includes IPDs or not: Yes or no	Authors
Longitude	Longitude of country of data collection	Authors
Latitude	Latitude of country of data collection	Authors
Region	Name of the region where the survey took place: Middle East & North Africa, Sub-Saharan Africa, South-Asia, Latin America & Caribbean	Authors
income_group	The income status of the country: Low income, Lower middle income, Upper middle income	Authors
country_of_birth	Binary indicator of whether the survey asks a respondent their country of birth: Yes or no	Authors
country_of_citizenship	Binary indicator of whether survey asks country of citizenship: Yes or no	Authors
acquisition_of_citizenship	Binary indicator of whether survey asks aquisition of citizenship: Yes or no	Authors
date_arrival	Binary indicator of whether the survey asks a respondent their date of arrival: Yes or no	Authors
reason_for_migration	Binary indicator of whether the survey asks their reason for migration: Yes or no	Authors

place_of_birth	Binary indicator of whether the survey asks a respondent their place of birth: Yes or no	Authors
date_of_first_dp	Binary indicator of whether the survey asks the date of first displacement: Yes or no	Authors
date_of_recent_dp	Binary indicator of whether the survey asks the date of recent displacement: Yes or no	Authors
main_reason_initial	Binary indicator of whether the survey asks the main reason of initial displacement: Yes or no	Authors
main_reason_recent	Binary indicator of whether the survey asks the main reason of most recent displacement: Yes or no	Authors
loc_habit	Binary indicator of whether the survey asks the location of habitual residence	Authors
loc_usual	Binary indicator of whether the survey asks the location of usual residence	Authors
tag_new_food_insecurity	Binary indicator of whether the dataset covers the topic of food insecurity	Authors
tag_new_health	Binary indicator of whether the dataset covers the topic of health	Authors
tag_new_water_sanitation	Binary indicator of whether the dataset covers the topic of water and sanitation	Authors
tag_new_education	Binary indicator of whether the dataset covers the topic of education	Authors
tag_new_welfare	Binary indicator of whether the dataset covers the topic of welfare	Authors
tag_new_labor_employment	Binary indicator of whether the dataset covers the topic of labor and employment	Authors
tag_new_finance_credit_debt	Binary indicator of whether the dataset covers the topic of finance, credit and debt	Authors
tag_new_protection	Binary indicator of whether the dataset covers the topic of protection	Authors
tag_new_gender	Binary indicator of whether the dataset covers the topic of gender	Authors
tag_new_safety_security_conflict	Binary indicator of whether the dataset covers the topic of safety and security	Authors
tag_new_housing_assets	Binary indicator of whether the dataset covers the topic of assets	Authors
tag_new_shocks	Binary indicator of whether the dataset covers the topic of shocks	Authors
tag_new_agriculture_livestock_fishery	Binary indicator of whether the dataset covers the topic of agriculture/livestock/fishery	Authors
tag_new_subjective_poverty	Binary indicator of whether the dataset covers the topic of subjective wellbeing	Authors
tag_new_coping_mechanisms	Binary indicator of whether the dataset covers the topic of coping mechanisms	Authors
tag_new_demographics	Binary indicator of whether the dataset covers the topic of demographics	Authors
tag_new_social	Binary indicator of whether the dataset covers the topic of social	Authors
tag_new_remmitance	Binary indicator of whether the dataset covers the topic of remittance	Authors
tag_new_child_labor	Binary indicator of whether the dataset covers the topic of child labor	Authors
tag_new_migration	Binary indicator of whether the dataset covers the topic of migration	Authors

Annex C: Taxonomy of Topic Areas

The following procedure is taken to tag the topic coverage of each micro-level dataset in the WB and UNHCR MDL. First, extract the information on the questionnaire modules in the metadata of the identified FDP datasets as well as all the non-FDP micro-level datasets in the WB MDL.[37] Second, calculate the frequency of words mentioned in those modules for both FDP datasets and non-FDP datasets. Third, map each frequently mentioned word to one of the 14 thematic areas. All words that are mentioned at least twice in FDP datasets and five times in non-FDP datasets[38] are manually mapped to these thematic areas. Table C 1 presents the list of keywords by topic areas. Finally, we use this information to tag each dataset by topic area based on whether its survey modules as described in the metadata mentioned any of the keywords listed in Table C 1. For example, if a given dataset includes a module containing at least one of the keywords listed under food insecurity (e.g., food, nutrition, malnutrition, stunting or nutritional), this dataset is tagged as relevant to the topic of "food insecurity".

Table C 1: Keywords and Thematic Categories

Topic	Words
Agriculture, livestock and fishery	agricultural, agriculture, agropecuaria, agrícolas, crop, crops, farmer, farmers, farming, fertilizer, fish, fishery, fishing, harvesting, harvests, irrigation, livestock, planting, plot, poultry
Assets	asset, assets, dwelling, dwellinghousing, dwellings, housing
Consumption and welfare	consumption, expenditure, income, nonfood, purchase, welfare
Coping mechanisms	coping, copingshocks
Education	classroom, educación, education, educational, educationliteracy, learning, literacy, maternelle, mathematics, numeracy, school, schooling, schools, teachers, teaching, éducation
Finance, credit, and debt	bank accounts, banks, borrow, borrowing, credit, debt, finance, finances, financial, lender, loans, saving, savings
Food (in)security	alimentation, dietary, food, food insecurity, food security, height, hungry, malnutrition, micronutrients, nutricion, nutrition, nutritional, nutritionnel, nutritious, stunting, underweight, vitamin, weight
Gender	domestic violence, female, femme, gender, gender violence, género, mujer, mujeres, woman, womans, women, womens
Health	aids, allaitement, anaemia, anemia, antenatal, anticoncepción, anticonceptivos, breastfeeding, cancer, clinical, clinician, contraception, contraceptive, contraceptives, contracepção, covid, covid19, covid19related, deworming, dhémoglobine, diabetes, diarrea, diarrhea, diarrhoea, diarrhée, disease, diseases, emergencyhealth, fecundidad, fecundidade, fertility, fertilitybirth, fièvre, fécondité, genital, haemoglobin, health, healthcare, healthy, hemoglobin, hemoglobina, hiv, hivaids, hospitalization, illness, immunisation, immunization, immunizations, infecciones, infections, inmunización, maladies, malaria, malariabednet, materna, medical, medicine, medicines, menstrual, morbidity, morbilidad, mortalidad, mortalidade, mortality, mortalité, mosquito, nets, paludisme, patients, postnatal, postnatals, pregnancy, prénatales, prénatals, reproduction, reproductiva, reproductivas, reproductive, santé, saúde, treatment, tuberculosis, vaccination, vaccinations, vaccine, vacunación, vih, vihsida
Labor and employment	business, businesses, earnings, empleo, emploi, employment, enterprise, enterprises, job, labor, labour, nonagricultural, nonfarm, pekerjaan, travail, unemployment, wage, wages, work, workers
Protection	protected, protection
Security and conflict	conflict, conflicts, safety, threats, violence, violencia

[37] For this analysis, we excluded those datasets that were either project specific or collected before 2010.

[38] These thresholds are determined to ensure that we cover at least 90 percent of all words mentioned in FDP and non-FDP datasets.

Shocks	shocks
Water and sanitation	defecation, drinking, handwashing, higiene, hygiene, latrine, nondrinking, sanitation, toilet, wash, water

Topic	Words
Agriculture, livestock and fishery	agricultural, agriculture, agropecuaria, agrícolas, crop, crops, farmer, farmers, farming, fertilizer, fish, fishery, fishing, harvesting, harvests, irrigation, livestock, planting, plot, poultry
Assets	asset, assets, dwelling, dwellinghousing, dwellings, housing
Consumption and welfare	consumption, expenditure, income, nonfood, purchase, welfare
Coping mechanisms	coping, copingshocks
Education	classroom, educación, education, educational, educationliteracy, learning, literacy, maternelle, mathematics, numeracy, school, schooling, schools, teachers, teaching, éducation
Finance, credit, and debt	bank accounts, banks, borrow, borrowing, credit, debt, finance, finances, financial, lender, loans, saving, savings
Food (in)security	alimentation, dietary, food, food insecurity, food security, height, hungry, malnutrition, micronutrients, nutricion, nutrition, nutritional, nutritionnel, nutritious, stunting, underweight, vitamin, weight
Gender	domestic violence, female, femme, gender, gender violence, género, mujer, mujeres, woman, womans, women, womens
Health	aids, allaitement, anaemia, anemia, antenatal, anticoncepción, anticonceptivos, breastfeeding, cancer, clinical, clinician, contraception, contraceptive, contraceptives, contracepção, covid, covid19, covid19related, deworming, dhémoglobine, diabetes, diarrea, diarrhea, diarrhoea, diarrhée, disease, diseases, emergencyhealth, fecundidad, fecundidade, fertility, fertilitybirth, fièvre, fécondité, genital, haemoglobin, health, healthcare, healthy, hemoglobin, hemoglobina, hiv, hivaids, hospitalization, illness, immunisation, immunization, immunizations, infecciones, infections, inmunización, maladies, malaria, malariabednet, materna, medical, medicine, medicines, menstrual, morbidity, morbilidad, mortalidad, mortalidade, mortality, mortalité, mosquito, nets, paludisme, patients, postnatal, postnatals, pregnancy, prénatales, prénatals, reproduction, reproductiva, reproductivas, reproductive, santé, saúde, treatment, tuberculosis, vaccination, vaccinations, vaccine, vacunación, vih, vihsida
Labor and employment	business, businesses, earnings, empleo, emploi, employment, enterprise, enterprises, job, labor, labour, nonagricultural, nonfarm, pekerjaan, travail, unemployment, wage, wages, work, workers
Protection	protected, protection
Security and conflict	conflict, conflicts, safety, threats, violence, violencia
Shocks	shocks
Water and sanitation	defecation, drinking, handwashing, higiene, hygiene, latrine, nondrinking, sanitation, toilet, wash, water

References

Aguilera, A., Krishnan, N., Muñoz, J., Riva, F.R., Sharma, D., Vishwanath, T. (2020). Sampling for Representative Surveys of Displaced Populations. In: Hoogeveen, J., Pape, U. (eds) Data Collection in Fragile States. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-25120-8_8.

Baal, N. 2021. "FORCED DISPLACEMENT DATA: Critical gaps and key opportunities in the context of the Global Compact on Refugees." Available at https://www.unhcr.org/people-forced-to-flee-book/wp-content/uploads/sites/137/2021/10/Natalia-Krynsky-Baal_Forced-Displacement-Data-Critical-gaps-and-key-opportunities-in-the-context-of-the-Global-Compact-on-Refugees.pdf.

Baal, N., and Ronkainen, L. 2017. Obtaining representative data on IDPs: challenges and recommendations. UNCHR Statistics Technical Series, Geneva, pp 1–9.

Berretta, M., Huang, C., Leon, M.D.A., and Lee, S. 2023. "Protocol: Addressing root causes and drivers of irregular migration – an Evidence Gap Map." Available at https://3ieimpact.org/sites/default/files/2023-05/IOM-Irregular-Migration-EGM-Protocol.pdf.

Campbell Collaboration. Campbell collaboration systematic reviews: policies and guidelines version 11 Oslo, Norway, 2015. Campbell Systematic Reviews. 2015. doi: 10.4073/csr.2015.1. http://www.campbellcollaboration.org/artman2/uploads/1/C2_Policies_and_Guidelines_Doc_Version_1_1-3.pdf Accessed 19 Oct 2015.Return to ref 9 in article.

Cas, S.C.M., Alem, Y., and Shirakawa, J.B. 2022. "Building Statistical Capacity in Fragile and Conflict-Affected States," IMF Working Papers 2022/045, International Monetary Fund.

Clapton J, Rutter D, Sharif N. SCIE Systematic mapping guidance; April 2009. http://www.scie.org.uk/publications/researchresources/rr03.pdf Accessed 19 Oct 2015.

Corral, P., Irwin, A., Krishnan, N., Mahler, D.G., and Vishwanath, T. 2020. *Fragility and Conflict: On the Front Lines of the Fight against Poverty*. Stand Alone Books.

Ginn, Thomas. 2023. "Labor Market Access and Outcomes for Refugees." World Bank - UNHCR Joint Data Center.

Gough D, Oliver S, Thomas J. An introduction to systematic reviews. London: Sage Publications Ltd; 2012.

James, K.L., Randall, N.P. and Haddaway, N.R. 2016. "A methodology for systematic mapping in environmental sciences." Environ Evid 5, 7 (2016). https://doi.org/10.1186/s13750-016-0059-6.

Rico, P., and Camilo, Juan. 2022. "Making migrants visible: a review of information on migrants in censuses and households surveys in Latin America and the Caribbean. Inter-American Development Bank, October 2022. http://dx.doi.org/10.18235/0004492.

Sabet, Shayda Mae and Annette N. Brown. 2018. "Is Impact Evaluation Still on the Rise? The new trends in 2010-2015." *Journal of Development Effectiveness.* no. 10, 3. 291—304. DOI: 10.1080/19439342.2018.1483414

Sida. 2020. “Migration and Development: Evidence Mapping Brief.” Available at https://www.dev-practitioners.eu/media/event-documents/2020_Sida_Evidence_mapping_brief_Migration.pdf.

Thompson, Kristi. “Data in Development: an Overview of Microdata on Developing Countries.” *IASSIST Quarterly*, Winter/Spring 2010. 25—30.

UNHCR. 2020. Global Report 2020: Populations of Concern to UNHCR.

USAID. 2021. “Mapping the Ecosystem of Education Data for Internally Displaced Persons in the Middle East and Beyond: Issues, Challenges, and Recommendations.” Available at https://www.edu-links.org/resources/mapping-ecosystem-education-data-internally-displaced-persons-middle-east-and-beyond.

European Union and United Nations (2018). *Expert Group on Refugee and Internally Displaced Persons Statistics: International Recommendations on Refugee Statistics*. Publications Office of the European Union: Luxembourg.

EGRISS. 2018. International Recommendations on Refugee Statistics (IRRS).

EGRISS. 2020. International Recommendations on Internally Displaced Persons Statistics (IRIS).

EGRISS. 2023. Compilers’ Manual on Forced Displacement Statistics. Use Case E. Sources of Operational Data from Humanitarian Organisations.

World Bank. 2021. World Development Report 2021: Data for Better Lives. World Bank.