Counter Trafficking Data Collaborative (CTDC)

The Global Synthetic Data Codebook

Version 2 – February 2024

1 DESCRIPTION OF DATA

The Global Synthetic Dataset, available for the first time in 2021 and currently updated in 2024, contains information on identified and reported victims of human trafficking. It comprises 27 variables on the socio-demographic profile of victims (such as gender or age), the trafficking process (such as means of control used on the victims), and the type of exploitation.

The Global Synthetic Dataset is available on the Counter Trafficking Data Collaborative (CTDC) website. The dataset combines contributions from all CTDC partners, including IOM, Polaris, RecollectiV (formerly Liberty Shared), A21, and the Portuguese Observatory on Trafficking in Human Beings (OTSH).[1] It represents assistance and hotline data from over 206,000 victims and survivors of trafficking identified across 190 countries and territories. It is the first and largest synthetic dataset of victims of trafficking case records.

2 DATA DE-IDENTIFICATION

Since July 2019, IOM has been working with Microsoft Research on a de-identification solution through the Tech Against Trafficking (TAT) accelerator program. Prior to the collaboration with Microsoft, the CTDC team used k-anonymization, which is another de-identification method, to protect the privacy and safety of victims and survivors. The limitation of the k-anonymization approach is that by redacting case records with rare combinations of quasi-identifiers, the overall size of the dataset can be reduced dramatically in ways that greatly distort data statistics. In the case of the original data, 40% of case records would need to be suppressed to protect the safety and privacy of survivors.[2]

With this latest approach, the dataset undergoes two stages of de-identification. In the first stage, all names and identifying details are removed from the data. In the second stage, the dataset is processed through a privacy-preserving pipeline. The main advantage of the Microsoft algorithm is that it overcomes the challenge of reduced sample size by synthesizing a new dataset in which records do not correspond to actual individuals, but which preserves the structure and statistics (i.e., utility) of the original data.

This is the third synthetic dataset derived from victims of trafficking case records with the guarantee of differential privacy. The Global Synthetic Victim-Perpetrator dataset is another differentially private dataset that is available for download. Differential privacy was developed at Microsoft Research in 2006, and today represents the gold standard in privacy protection. The idea is that if you get a similar answer to any data query whether any individual data subject is in the dataset used to answer the query, then you cannot infer the presence of that individual in the dataset. This is

[1] A21 and OTSH were not able to contribute their latest data at the time of this data update in 2024, but they plan to contribute to the next data update. RecollectiV discontinued their data contribution since 2020 due to changes in work modality.

[2] The original data implies the sensitive input data that are used to generate the synthetic data in the codebook.

true no matter your background knowledge, including knowing the answers to earlier queries, or how many subjects have been added to the dataset since those queries. In short, it addresses the theoretical privacy gap left by k-anonymity whenever there is an expectation of multiple overlapping data releases over time.

More information on the approach is available through the open-source software and documentation on differential privacy via GitHub. Please refer to **Appendix I: Microsoft Research's Approach to Generating Synthetic Data with Differential Privacy**.

3 DATA SOURCE

The data included in the Global Synthetic Dataset is the result of an extensive process of comparing contributors' existing data classification systems, to identify data definitions which were identical or compatible. As CTDC co-founders, Polaris and IOM initially agreed on a shared lexicon and format. Where possible, CTDC follows international standard classification. Some concepts follow operational definitions as there are no internationally agreed definitions, such as types of sexual exploitation.

In 2023, IOM and UNODC published the working version of the International Classification Standard for Administrative Data on Trafficking in Persons (ICS-TIP) with the aim to harness safe, effective, and standardized administrative data. Even though CTDC contributors' data are collected before the inception of the ICS-TIP, the units of description and classification (i.e., event, victim, perpetrator, reporting entity) and variable definitions of the CTDC data are largely in line with the ICS-TIP.

3.1 IOM

IOM's data collection processes have evolved over time. IOM has recorded case data on victims of trafficking through assistance programmes since 2002. Records from 2002 to 2011 are called "legacy records". In 2012, IOM rolled out a web-based case management system called the "Migrant Management Operational System Application" (MiMOSA); IOM's missions gradually moved to MiMOSA to record identified and assisted victims of trafficking. By 2014, all missions switched to entering case data using the MiMOSA web form. MiMOSA has many data fields – from socio-demographic to route data. By 2019, MiMOSA has been upgraded and some questions in the web form have been updated. IOM's data have been cross-referenced across different databases in-house and over time.

The type of data collected reflects the organization's operations. IOM offers comprehensive direct assistance packages to victims of trafficking in the context of its programmes and collaboration with governments and civil society partners. The extent to which IOM has direct contact with victims and for which purposes vary. In some contexts, IOM identifies victims of trafficking and provides them with immediate assistance. In some countries, IOM effectively acts as the national referral mechanism for all identified victims of trafficking. In cases where victims have been trafficked out of their home countries, IOM could be involved in facilitating their safe, voluntary return to their countries of origin and ensuring long-term assistance and re-integration, sometimes through referral to implementing partners in their countries of origin. In some cases, victims of trafficking would only be referred to IOM after they returned to their countries of origin. As the leading international organization providing direct assistance to victims of trafficking, IOM is involved in bridging different national contexts and facilitating the process of returning to home countries or

resettlement to a third country. More information on IOM's direct assistance activities and data is available here.

3.2 Polaris

Polaris' dataset is comprised of information obtained during Polaris' regular interactions with individuals contacting the U.S. National Human Trafficking Hotline and the BeFree Textline, both operated by Polaris. Polaris does not investigate claims made by individuals contacting the helplines and cannot verify the accuracy of the information reported. This data is not the result of a systematic survey. As these individuals told their own stories or relayed the experiences of their friends and family members, Polaris staff noted key elements of each account. This information was later classified in over 120 standardized fields using detailed standards and definitions and this information constitutes the dataset contributed by Polaris. Victims and third parties reporting these situations were not asked a set of standardized questions and only provided information that they felt comfortable sharing with Polaris's staff to get the help they needed. Upon request, Polaris removes information about individuals who do not wish to be included in the dataset.

Polaris has operated the National Human Trafficking Hotline since December 7, 2007 and the BeFree Textline since March 28, 2013. At present, Polaris is only able to contribute victim data from cases reported to the National Human Trafficking Hotline and the BeFree Textline since January 1, 2015 as the structure of data collected prior to this date is not compatible with CTDC's data standards. Polaris has reclassified its historical data and has been contributing data reported after March 31, 2017 on a biannual basis. More information about Polaris and its data is available here.

3.3 RecollectiV (formerly Liberty Shared)

RecollectiV has discontinued their contribution to CTDC from the 2024 update due to changes in their work modality with counter-trafficking NGOs, as they will no longer have access to NGOs' case management data. Although CTDC has lost one data contributor, RecollectiV's historical data will remain in the original data as it is being updated.

RecollectiV's data have been collected through their Victim Case Management System (VCMS). The VCMS is a cloud-based data collection and information management tool, designed by Liberty Shared to assist frontline NGOs in combating human trafficking and modern slavery through robust record-keeping. The VCMS brings together over 40 NGO partners working on this common challenge onto a shared platform. It facilitates standardized data collection.

The VCMS has many data fields available to assist frontline NGO users in recording information related to victims and cases. Data points cover the critical components of a victim's experience, from pre-exploitation demographics, through the recruitment, transit, and exploitation phases. NGO VCMS users receive training and ongoing support from RecollectiV to record information.

The data contributed to the Global Synthetic Dataset by RecollectiV comes directly from NGO partners utilizing the VCMS who have consented to be a part of CTDC. Some of this data has been directly entered by NGO users into the VCMS, which has been operational since 2014, whilst some data contributed by these partners is "legacy" data, migrated into the VCMS from their existing data storage setups when they transitioned onto the VCMS platform. More information on RecollectiV and their VCMS is available here.

3.4 A21

A21 is a global non-governmental organization combating human trafficking through prevention and awareness, intervention, and aftercare. A21's data provided to CTDC consists of information garnered from case files dating back to 2008 from their seven aftercare countries – Bulgaria, Cambodia, Greece, South Africa, Thailand, Ukraine, and the United States.

As A21 is a frontline service provider, all data is a result of direct engagement with survivors from the point of their identification and recovery through to independence. A21 continuously collects and updates data at intake and from regular interactions with survivors through service provisions. The data reveal a robust picture of survivors' unique trafficking situations and factors of vulnerability. More information on A21 and their data is available here.

3.5 OTSH

OTSH was established (Decree-law no. 229/2008) in response to the opacity which characterises the trafficking in human beings phenomenon and thus, through enhanced understanding, contributes to better forms of intervention regarding prevention, protection and prosecution. The Observatory is part of the Ministry of Home Affairs of Portugal and develops its work in close cooperation with the National Coordinator. The mission of the Observatory is to produce, collect, analyse and disseminate information and knowledge about trafficking in human beings and other forms of gender violence. More information on OTSH and their data is available here.

4 DATA LIMITATIONS

Randomness: Data are only available where the contributing organizations are operational and can share such data. Therefore, this dataset does not include a random sample of trafficking victims worldwide. Nevertheless, in the countries where IOM provides direct assistance to victims of trafficking, the data can be considered broadly representative of the identified victim population in the country.

Consistency: CTDC partners comprise different organizations, with different data collection methods, different types of assistance, as well as follow-up services. The CTDC team shares some resources, e.g., data dictionary and data preparation instructions, to ensure the data are coded in a consistent manner.

Bias: Potential victims of human trafficking need to be either identified or (self-) reported. This sample may be biased if certain types of trafficking or socio-demographic groups are more likely to be identified or referred to than others. Since the unidentified population is by definition unknown, the extent of bias is not known and cannot be corrected for.

5 LIST OF VARIABLES

1 yearOfRegistration
2 gender
3 ageBroad
4 citizenship
5 CountryOfExploitation
6 traffickMonths
7 meansDebtBondageEarnings
8 meansThreats
9 meansAbusePsyPhySex
10 meansFalsePromises
11 meansDrugsAlcohol
12 meansDenyBasicNeeds
13 meansExcessiveWorkHours
14 meansWithholdDocs
15 isForcedLabour
16 isSexualExploit
17 isOtherExploit
18 typeOfLabourAgriculture
19 typeOfLabourConstruction
20 typeOfLabourDomesticWork
21 typeOfLabourHospitality
22 typeOfSexProstitution
23 typeOfSexPornography
24 recruiterRelationIntimatePartner
25 recruiterRelationFriend
26 recruiterRelationFamily
27 recruiterRelationOther

Notes: Some of the means of control variables from the previous version of Global Synthetic Data are merged following Stockl et al. (2021). The following variables are in the previous version of the Global Synthetic Data but are no longer exported: *meansOfControlOther, typeOfLabourOther,* and *typeOfSexOther*. Please refer to **Appendix II** for further information on the changes.

6 PERCENTAGE CALCULATIONS

To calculate the percentages of means of control (variables 7-14), type of exploitation (variables 15-17), type of labour exploitation (variables 18-21), type of sexual exploitation (variables 22-23), and/or trafficked person's relationship with the recruiter (variables 24-27), one needs to define the denominator.

Here is an example of how the denominator is defined by type of exploitation at the case level. When the trafficked person provides *any* information on the type of exploitation, it represents the number 1. Then, the other types of exploitation should be coded as 0 as the denominator.

Example: Defining the denominator by type of exploitation

Synthetic data

Person	isForcedLabour	isSexualExploit	isOtherExploit
1	1		
2		1	
3			1
4			
5	1	1	

Processed data

Person	isForcedLabour	isSexualExploit	isOtherExploit
1	1	0	0
2	0	1	0
3	0	0	1
4			
5	1	1	0

% by type of exploitation	50	50	25

If Person 4 (in the Example) would be counted, the share of persons in trafficking for the purpose of forced labour and sexual exploitation would be 40% each (2/5*100%=40%; whereas the share of those in other types of exploitation would be 20% (1/5*100%=20%). The CTDC team recommends including *only* those who provided information on the type of exploitation in calculating the denominator, to gain a more accurate idea of the share of each type of trafficking among identified victims.

As an individual can experience multiple types of exploitation (or means of control, types of labour/sexual exploitation, relationships with the recruiter), the sum of each group does not necessarily add up to 100%. In this example, % *isForcedLabour* + % *isSexualExploit* + % *isOtherExploit* = 125%.

Note: The percentages are determined by the proportion of cases in which the attribute value was observed. While the total attribute counts and relative magnitudes are preserved, the calculated percentages from the synthetic data may vary compared to the original data.

If you will report percentages in any public-facing documents, please clearly state the source of the data (i.e., Source: Counter-Trafficking Data Collaborative (CTDC). 2024. 'Global Synthetic Dataset'. Available at: https://www.ctdatacollaborative.org/page/global-synthetic-dataset (Accessed Day Month Year)).

7 DETAILED DESCRIPTION OF VARIABLES

Before undertaking exploratory analysis, please consider that a significant number of missing values for a variable may be a result of the fact that only certain CTDC contributors have data collected.

Variable 1

Variable name: yearOfRegistration
Variable label: Year of Registration
Type: numeric
Values and categories:

- Range: [2002, 2022]

Definition: The year in which the individual was assisted/identified/referred/reported to the contributing organizations.

Variable 2

Variable name: gender
Variable label: Gender
Type: string
Values and categories:

- NULL [*missing data*]
- Man [*The individual identifies with a gender role most usually attributed to a man by relevant culture and society*]
- Woman [*The individual identifies with a gender role most usually attributed to a woman by relevant culture and society*]
- Trans/Transgender/NonConforming [*The individuals' expression of gender falls outside binary societal expectations of what a man and a woman are. Including, but not limited to: Individuals whose gender identity differs from what is typically associated with the sex they were assigned at birth or falls outside the male-female binary. For data privacy reasons, these data are not further disaggregated in this dataset.*]

Since 2002, the understanding and approach to gender considerations have evolved considerably. In the older versions of the data, and some contexts, gender and sex (i.e., sex assigned at birth) are not differentiated. For example, IOM has started collecting data about gender which includes trans/transgender categories since 2017.

Definition: Gender is a social designation (rather than sex, which is a classification usually assigned at birth), referring to men, women, and people with diverse genders. Even though CTDC partners collect data on other gender identities that consider the psychological, behavioural, social, and cultural aspects of being trans/transgender/non-conforming, or not specified/unknown, the number of observations is too small to be described in detail in this dataset.

Variable 3

Variable name: ageBroad
Variable label: Age Broad
Type: string
Values and categories:

- NULL [*missing data*]
- 0—8
- 9—17
- 18—20
- 21—23
- 24—26
- 27—29
- 30—38
- 39—47
- 48+

Definition: The individual's age at the time the individual was assisted/identified/referred/reported to the contributing organizations.

Variable 4

Variable name: citizenship
Variable label: Citizenship
Type: string
Values and categories:

- NULL *[missing data]*
- Values based on ISO 3166-1 Alpha-3 Codes

Definition: The set of rights and duties that a person has with a country because of his/her legal bond with the country. This term is often used interchangeably with nationality. RecollectiV records this as 'Country of Origin' as a proxy for citizenship. Please refer to **Appendix II** for more information. To protect the privacy of individuals, some countries of citizenship are suppressed as they are highly sensitive and cannot be protected.

Variable 5

Variable name: CountryOfExploitation
Variable label: Country of Exploitation
Type: string
Values and categories:

- NULL *[missing data]*
- Values based on ISO 3166-1 Alpha-3 Codes

Definition: This variable indicates the country where a victim is first supported/assisted, identified and/or referred. In the context of human trafficking data, this is also referred to as the "country of destination2" (as opposed to "country of origin" if human trafficking was across borders). In the case of IOM, the last country of exploitation has a high number of missing values. Therefore, a proxy has been created to capture as much data as possible about the country the victim was exploited in. Please refer to **Appendix II** for more information. To protect the privacy of individuals, some countries of exploitation are suppressed as they are highly sensitive and cannot be protected.

Variable 6

Variable name: traffickMonths
Variable label: Trafficking Duration, in Months
Type: string
Values and categories:

- NULL *[missing data]*
- 0—12 (0-1 yr)
- 13—24 (1-2 yrs)
- 25+ (2+ yrs)

Definition: Indicates the reported duration of trafficking in months.

Variable 7

Variable name: meansDebtBondageEarnings
Variable label: Means of Control: Debt Bondage and/or Withhold Wages
Type: binary numeric
Values and categories:

- NULL *[missing data]*
- 1

Definition: To improve the synthetic data generation, this variable merges two possible means of control: debt bondage and take earnings. The variable indicates whether the individual is forced to work to pay off a created or perceived debt and/or whether the individual has experienced a situation where the exploiters have taken his/her remuneration. The individual is deceived to work for little or no pay, with no control over his/her debt. Debt bondage is defined as the status or condition arising from a pledge by a debtor of his personal services or those of a person under his control as security for a debt, if the value of those services as reasonably assessed is not applied towards the liquidation of the debt or the length and nature of those services are not respectively limited and defined (United Nations' 1956 Supplementary Convention on the Abolition of Slavery).

Variable 8

Variable name: meansThreats
Variable label: Means of Control: Threats to Individual and/or Family
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: To improve the synthetic data generation, this variable merges three possible means of control: threats, use of children, and threat of law enforcement. The variable indicates whether the individual experienced a situation in which his/her exploiter(s) explicitly or implicitly communicated an intent to inflict harm or loss on the individual or others, such as their family members. It also indicates whether the individual experienced a situation in which their exploiter(s) explicitly or implicitly communicated an intent to contact or involve law enforcement or other relevant authorities, such as immigration authorities, to negatively impact the individual or others, such as their family members.

Variable 9

Variable name: meansAbusePsyPhySex
Variable label: Means of Control: Abuse (Psychological, Physical, and/or Sexual)
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: To improve the synthetic data generation, this variable merges three possible means of control: psychological abuse, physical abuse, and sexual abuse.

'Psychological abuse' indicates whether the individual experienced a situation in which their exploiter(s) used various forms of abuse. This includes psychological abuse through emotionally abusive, deceptive, or devious tactics to influence the individual. This may include but is not limited to, name-calling, verbal abuse, humiliating in front of others, manipulating real or perceived power imbalances, shaming, or blaming the individual. It may also include acts intended to exploit or prey upon any familial or romantic bonds/attachments the individual has with their exploiter(s).

'Physical abuse' indicates whether the individual experienced a situation of physical abuse involving actions to cause physical injury, pain, disability, death or trauma to the individual. This includes but is not limited to shoving, strangulation, shaking, slapping, punching, kicking, pulling hair, burning, branding or tattooing, the use of a weapon, or using one's size and strength against the individual.

'Sexual abuse' indicates whether the individual experienced sexual abuse encompassing any kind of unwanted or non-consenting sexual contact from their exploiter(s) as a means by which to control the individual, rather than a purpose for which the individual was trafficked, in order to influence their behaviour. This includes but is not limited to, using sexual acts, assault, or contact as punishment, manipulation, or normalizing sexual violence. It also includes coercive behaviour that interferes with the individual's ability to control his/her reproductive life, including but not limited to, forcing/coercing the individual to terminate or continue a pregnancy against their will, manipulating birth control, intentionally exposing someone to an STI, preventing condom negotiation, and/or attempting to or impregnating the individual without their consent.

Variable 10

Variable name: meansFalsePromises
Variable label: Means of Control: False promises
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual was defrauded or tricked into entering the exploitative situation by their exploiter(s) using deception and false pretences.

Variable 11

Variable name: meansDrugsAlcohol
Variable label: Means of Control: Psychoactive Substances
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the exploiter(s) induced the individual into substance abuse, provided substances to make the individual compliant or in order to influence their behaviour, or exploited an existing substance abuse issue.

Variable 12

Variable name: meansDenyBasicNeeds
Variable label: Means of Control: Restricts Finance, Movement, Medical Care, and/or Necessities
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: To improve the synthetic data generation, this variable merges four possible means of control: restrict financial access, restrict movement, restrict medical care, and withhold necessities.

‘Restrict financial access’ indicates whether the individual experienced a situation in which his/her exploiter(s) prohibited or restricted the individual's access to necessary daily living funds or their own personal finances. This includes but is not limited to, controlling an individual’s personal bank account, bank/credit cards, or overtly stealing the individual’s personal funds.

‘Restrict movement’ indicates whether the exploiter(s) isolated, confined or limited the movement of the individual in any way physically or socially. This may include situations in which the individual is physically detained, prevented from moving without being accompanied or monitored, or the exploiter(s) threatens or enacts repercussions related to the individual’s movement. This may also include forms of emotional isolation including restricting the individual’s access to support systems or social networks or moving the individual frequently to prevent the individual from establishing support systems or social networks.

'Restrict medical care’ indicates whether the exploiter(s) limited the individual’s access to medical or health services. This includes but is not limited to necessary medical care or treatment being

withheld, or when access to such treatment was controlled by the exploiter(s). This also includes situations in which the individual was unable to access or interact with health services without being accompanied or monitored by the exploiter(s).

'Withhold necessities' indicates whether the individual experienced a situation in which their exploiter(s) denied, restricted, or threatened to deny/restrict basic living necessities such as food, shelter, water, hygiene, appropriate clothes, or necessary items for religious observance or gender expression.

Variable 13

Variable name: meansExcessiveWorkHours
Variable label: Means of Control: Excessive Working Hours
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual was required to work a significant number of hours more than what they were contracted or promised. It could include overtime, late or atypical shifts, or overnight hours. It could also be used as a means of keeping the individual isolated and/or unable to seek help or report their situation. In some instances, work hours may be so excessive as to cause physical and/or mental health issues which may also limit the individual's capacity to seek help or report their situation.

Variable 14

Variable name: meansWithholdDocuments
Variable label: Means of Control: Withhold Documents
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced a situation in which their exploiter(s) limited, restricted, or controlled the individual's access to important documents including, but not limited to, the individual's passport, immigration documents, work permit, identification card, government benefit documents, birth certificate, gender identity carry letter, court-issued protection orders, custody papers, or other legal, official, or government documents.

Variable 15

Variable name: isForcedLabour
Variable label: Type of Trafficking: Forced Labour
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: It indicates that the purpose for which a victim was trafficked was all work or service which was exacted from the individual, under the threat of a penalty and for which the individual has not offered himself or herself voluntarily. Sexual services are excluded from this definition.

Variable 16

Variable name: isSexualExploit
Variable label: Type of Trafficking: Sexual Exploitation
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: It indicates that the purpose for which a victim was trafficked was sexual services, such as the exploitation of the prostitution of an individual. Typically, the exploiter(s) use force, fraud or coercion to achieve exploitation.

Variable 17

Variable name: isOtherExploit
Variable label: Type of Trafficking: Other Exploit
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: It indicates that the purpose for which a victim was trafficked was other types of exploitation, including those who underwent forced marriage, forced military, and organ removal. Typically, the exploiter(s) use force, fraud, or coercion to achieve exploitation. Polaris staff apply the U.S. federal definition of human trafficking as defined by the Trafficking Victims Protection Action (TVPA) to determine if a situation described through the helplines has indications of human trafficking. Situations of forced marriage, organ harvesting or “other”, which do not meet the U.S. definition of trafficking are not included in the data contributed by Polaris to the CTDC as Polaris currently does not collect data about these sub-types.

Variable 18

Variable name: typeOfLabourAgriculture
Variable label: Type of Labour Exploit: Agriculture
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced forced labour in activities defined in ISIC Section A Division 01, "Crop and animal production, hunting and related service activities". This category does not include work related to forestry and logging as defined by ISIC Section A, Division 02 which is considered distinct.

Variable 19

Variable name: typeOfLabourConstruction
Variable label: Type of Labour Exploit: Construction
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced forced labour in activities defined in ISIC Section F "Construction".

Variable 20

Variable name: typeOfLabourDomesticWork
Variable label: Type of Labour Exploit: Domestic Work
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced forced labour in activities defined in ISIC Section T "Activities of households as employers; undifferentiated goods- and services- producing activities of households for own use".

Variable 21

Variable name: typeOfLabourHospitality
Variable label: Type of Labour Exploit: Hospitality
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced forced labour in activities defined in ISIC Section I "Accommodation and food service activities" including both Division 55 "Accommodation" and Division 56 "Food and beverage service activities".

Variable 22

Variable name: typeOfSexProstitution
Variable label: Type of Sexual Exploit: Prostitution
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced exploitation in activities associated with an individual's sexual act for payment.

Variable 23

Variable name: typeOfSexPornography
Variable label: Type of Sexual Exploit: Pornography
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether the individual experienced exploitation in the production of visual material depicting sexual behaviour that is intended to arouse sexual excitement in its audience and does not involve any participation from the audience. It does not include situations in which the audience remotely participates or interacts with the individual featured in the visual material.

Variable 24

Variable name: recruiterRelationIntimatePartner
Variable label: Recruiter Relation: Intimate Partner
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether a person who initially enticed or obtained the individual into the situation of exploitation was one with whom the individual has identified as having a current or former romantic relationship.

Variable 25

Variable name: recruiterRelationFriend
Variable label: Recruiter Relation: Friend
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether a person who initially enticed or obtained the individual into the situation of exploitation was one with whom the individual was familiar, exclusive of romantic partners, family relations, or other more formal relationships.

Variable 26

Variable name: recruiterRelationFamily
Variable label: Recruiter Relation: Family
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether a person who initially enticed or obtained the individual into the situation of exploitation was one with whom the individual was connected biologically, through marriage, custodianship, or guardianship. This may include but is not limited to, parents, primary caregivers, and foster parents.

Variable 27

Variable name: recruiterRelationOther
Variable label: Recruiter Relation: Other
Type: binary numeric
Values and categories:

- NULL [*missing data*]
- 1

Definition: Indicates whether a person who initially enticed or obtained the individual into the situation of exploitation was a person with whom the individual had any other notable relationship that cannot reasonably fit into previous categories. This may include, but is not limited to, labour brokers, contractors, formal employers, or smugglers. In the case of RecollectiV, data collected on recruiters was more extensive than Global Dataset required. RecollectiV transformed values to match IOM values, resulting in a higher level of 'other' values where information recorded did not match (for example – Broker – becomes other).

Appendix I: Microsoft Research's Approach to Generating Synthetic Data with Differential Privacy

One aspect of victim safety is ensuring the privacy of data subjects.[3] Such privacy means that traffickers are prevented from identifying known victims in published datasets, making those victims safe from reprisals. Another aspect of victim safety is ensuring the accuracy of data statistics. Such accuracy means that downstream activities of data-driven decision-making and policy-making are based on the best available data, leading to the most appropriate actions.

The challenge for safe data sharing is that the methods used to preserve the privacy of data subjects typically distort data statistics, and if they are distorted in the wrong ways then this could lead to misguided actions that compromise the safety of the broader victim population. For example, if a privacy method greatly over- or under-reported a given case pattern – or fabricated it entirely – this could mislead decision-makers into misallocating scarce resources in ways that fail to tackle the actual problems observed.

Microsoft Research's approaches are based on the idea that rather than redacting sensitive data to create privacy, one can instead generate synthetic data that is private by design, yet accurately captures the structure and statistics of the underlying sensitive dataset.

In the September 2021 release of the Global Synthetic Dataset, IOM used a new algorithm from Microsoft Research that generated synthetic data with k-anonymity for all combinations of attributes, not just the subset of attributes labelled in advance as quasi-identifiers. This addressed both the privacy and accuracy limitations of the k-anonymization method and earlier release – all combinations of attributes included in the synthetic and aggregate datasets appeared at least k times in the original sensitive dataset and were reported precisely in the aggregate dataset (rounded down to the closest k).

While synthetic data with "full" k-anonymity represented a major improvement over standard k-anonymization, the nature of the privacy guarantee only extends to a single release. Across a series of multiple releases (where each version builds on the data in the previous release), k-anonymity provides no theoretical guarantees about what an attacker could potentially learn by comparing the differences between reported counts and the records they know (or assume) have been added to the dataset. Although a large enough k and a sufficient record increment are likely to guard against all practical attacks, there is another privacy paradigm that is explicitly designed to combat such "differencing" attacks: differential privacy.

Differential privacy was developed at Microsoft Research in 2006, and today represents the gold standard in privacy protection. The idea is that if one gets a similar answer to any data query, whether or not any individual data subject is in the dataset used to answer the query, then one cannot infer the presence of that individual in the dataset. This is true no matter one's background knowledge, including the answers to earlier queries, or how many subjects

[3] The CTDC team would like to thank Darren Edge and Rodrigo Racanicci (Microsoft Research) for their time in explaining the differential privacy approach and the synthetic data generation process to the team.

have been added to the dataset since those queries. In short, differential privacy addresses the theoretical privacy gap left by k-anonymity whenever there is an expectation of multiple overlapping data releases over time.

A central concept in differential privacy is the idea of quantifiable privacy loss – the extent to which the answers to arbitrary data queries are allowed to vary, probabilistically, based on the presence or absence of individual data subjects. The parameter that captures this concept is called epsilon. It serves as a budget for the allowable privacy loss across all queries. Each time the sensitive data is queried, calibrated noise is added to the answer in ways that control the possible privacy loss, and part of this budget is consumed. Once the budget is exhausted, no more queries can be answered.

To create synthetic data with differential privacy, the new method from Microsoft Research first uses the privacy budget to query the counts of cases matching all short combinations of case attributes. The results of these queries are released as aggregate data with differential privacy. Synthetic records are then constructed by sampling these combinations based on their noisy counts until all attributes in the sensitive dataset (based on the noisy counts in the aggregate dataset) have been accounted for. The resulting synthetic dataset retains the same degree of differential privacy as the aggregate data used as an input, and the worst-case privacy loss across a series of releases is simply the sum of the individual privacy budgets used to generate them. The release of the synthetic case records and aggregate data enables the evaluation of synthetic data accuracy as well as the retrieval of accurate counts (e.g., for official reporting).

For more details on the overall approach, including proof of differential privacy, see https://github.com/microsoft/synthetic-data-showcase.

Appendix II: Additional considerations on variables in the Global Synthetic Dataset

- General comment on variables 4 (citizenship) and 5 (CountryOfExploitation)
    - Serbia: includes Kosovo following UN guidelines.
    - China: includes Tibet.

- Country of exploitation proxy variable (Variable 5 – CountryOfExploitation)

  This is usually the country where a victim is first identified and/or referred to IOM. This does not necessarily capture the last part of the victim's journey before they are identified and referred nor the intended country of destination, which is not currently captured in the IOM database. The data points that are recorded in the case management system are 'last country of exploitation', 'location of the screening interview', 'location of the victim upon registration', and 'location of the IOM mission registering the case'.

- Some means of control variables are merged or renamed to align the naming format

	Variable name (2024 version)	**Variable name (2021 / previous version)**
7	meansDebtBondageEarnings	Merge (meansOfControlDebtBondage; meansOfControlTakesEarnings)
8	meansThreats	Merge (meansOfControlThreats; meansOfControlUsesChildren; meansOfControlThreatOfLawEnforcement)
9	meansAbusePsyPhySex	Merge (meansOfControlPsychologicalAbuse; meansOfControlPhysicalAbuse; meansOfControlSexualAbuse)
10	meansFalsePromises	Align naming format (meansOfControlFalsePromises)
11	meansDrugsAlcohol	Align naming format (meansOfControlPsychoactiveSubstances)
12	meansDenyBasicNeeds	Merge (meansOfControlRestrictsFinancialAccess; meansOfControlRestrictsMovement; meansOfControlRestrictsMedicalCare; meansOfControlWithholdsNecessities)
13	meansExcessiveWorkHours	Align naming format (meansOfControlExcessiveWorkingHours)
14	meansWithholdDocs	Align naming format (meansOfControlWithholdsDocuments)