

Synthetic Data

IOM/Microsoft collaboration

Prepared by the Counter-Trafficking Data Collaborative (CTDC)

2025-02-26



What are synthetic data?

- Synthetic data are **artificial data** generated from real data
- It cannot be linked back to the original data (including cases), so we can ensure data confidentiality and privacy.
- The statistical properties and relationships from the original data are preserved in synthetic data.
- The advantages of synthetic data over other forms of data are, e.g.,
 - **Aggregate data:** Researchers can no longer make connections between traits
 - Example: The Global Estimates of Modern Slavery
 - **K-anonymized data:** The redaction of outliers can lead to significant data loss
 - Example: CTDC data pre-2021

Example: Raw Data vs. K-anonymized Data

Raw Data

Anonymization

K-anonymized Data ($k = 2$)

Gender	Age	isForcedLabour
Female	19	1
Male	18	1
Male	20	1
Male	37	
Female	35	
Female	31	

Gender	AgeBroad	isForcedLabour	k
Female	18-20	1	1
Male	18-20	1	2
Male	18-20	1	2
Male	30-38		1
Female	30-38		2
Female	30-38		2

Example: Synthetic Data vs. K-anonymized Data

Synthetic Data (?)

Gender	AgeBroad	isForcedLabour
Female		
	18–20	1
Male		
	30–38	
		1
Male	18–20	1
Male	18–20	1
Female	30–38	
Female	30–38	

K-anonymized Data ($k = 2$)

Gender	AgeBroad	isForcedLabour	k
Female	18–20	1	1
Male	18–20	1	2
Male	18–20	1	2
Male	30–38		1
Female	30–38		2
Female	30–38		2

**Both datasets are generated from the same raw data. For CTDC, $k = 10$.*

Overcoming challenges with synthetic data

Problem – Then

- CTDC's previous solution was **labour-intensive and partner reliant**
- K-anonymization results in the **loss** of potentially useful data
- Risk of **re-identification and reprisals**
- Market providers were **expensive**

Solution – Now

- Microsoft Research's solution **automatically** prevents the publication of rare attributes
- **Preserves** the statistical properties and relationships in the original data
- Differential privacy **guarantees against any privacy attacks**
- **Open-source**

Synthetic Data Timeline and achievements



Synthetic Data

Tech Against Trafficking Accelerator Program

- Tech Against Trafficking invited IOM's CTDC to participate the 2019 Accelerator Program.
- The partnership focused on 3 workstreams:
 1. **Privacy-preserving mechanism:**
Develop a solution for analyzing case data while protecting victim privacy.
 2. **Data standards:**
Address data standards/consistency related to victim case management.
 3. **Stakeholder engagement:**
Maximize utility and impact of the CTDC platform.
- IOM has benefitted from substantial in-kind support from Microsoft Research to support CTDC.

Microsoft Research's Intelligence Toolkit

- AI-powered suite for large-scale data analysis, featuring case, entity, and text intelligence workflows.
- The Anonymise Case Data workflow generates synthetic data, aggregate data, and data visualization and reports.
- Accessible via web app (Azure or Docker) or Python package (PyPI).

Prepare

Generate

Query and Visualise

Intelligence Toolkit

Free-to-use AI-powered suite

[Anonymize Case Data workflow](#): [Prepare sensitive data](#) [Generate anonymous data](#) [Query and visualize data](#) [View example outputs](#)

Anonymize Case Data

The [Anonymize Case Data](#) workflow generates differentially-private datasets and data summaries from sensitive case records.

Select the [View example outputs](#) tab (in app) or navigate to [example_outputs/anonymize_case_data](#) (on GitHub) for examples.

Input requirements

- The input data file should be in CSV format with each row representing a different case (i.e., individual person or data subject).
- For data synthesis, each case must be represented as a collection of discrete (i.e., categorical or binary) attributes. Any continuous attributes must first be quantized via the user interface.
- Given the goal of creating an anonymous dataset, no direct identifiers (e.g., names, aliases, ids, phone numbers, email addresses, street addresses) should be included in data outputs. Following the principle of [data minimization](#), such direct identifiers should be removed from data inputs because they are not required for the processing purpose and create unnecessary risks for the data subject. Tools such as Microsoft Excel can be used to delete any direct identifier columns prior to use in Intelligence Toolkit.
- The nature of differential privacy means that indirect identifiers (e.g., age range, year of birth, gender, country, city) may be freely included in the data inputs. None of the combinations of these identifiers (or of any attributes) in the output data allow the presence of individuals to be inferred with any degree of certainty.

Use with other workflows

[Anonymize Case Data](#) can be used to anonymize case data for privacy-preserving analysis in any other workflow accepting structured records as input:

Global Synthetic Dataset (2021)

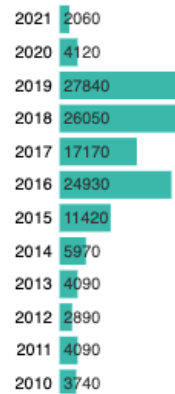
You can download data, codebook, and data dictionary

CTDC Global Dataset on Victims of Trafficking

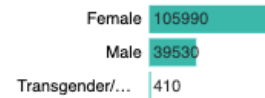
Privacy resolution (10): the minimum group size detectable in synthetic/aggregate data
Estimated counts: from synthetic data that reflects the sensitive data at the given resolution
Actual counts: from aggregate data rounded down to the closest multiple of the resolution



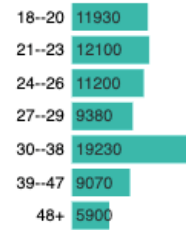
yearOfRegistration



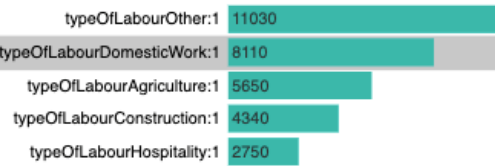
gender



ageBroad



Type of Labour Exploit



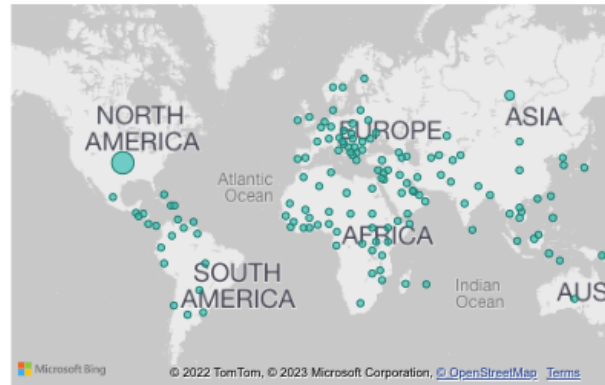
Type of Sexual Exploit



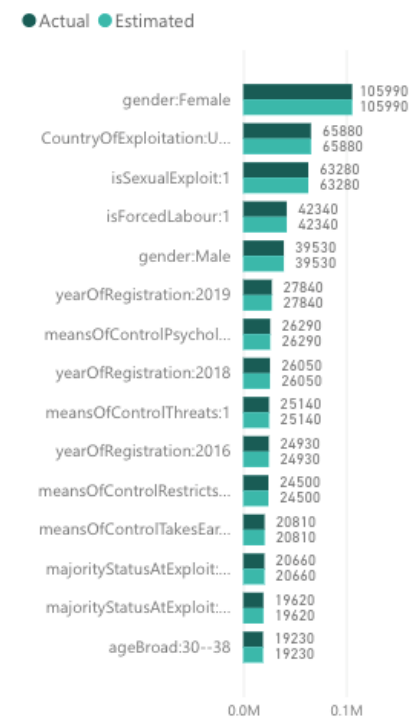
citizenship



CountryOfExploitation



Comparison for Selections



Victim's characteristics, e.g.,

- Country of exploitation
- Citizenship
- Type of trafficking, labour exploit, sexual exploit
- Means of control
- Trafficking duration
- and more...

Global Victim-Perpetrator Synthetic Dataset (2023)

You can download data, codebook, and data dictionary

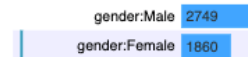
Privacy-preserving data on victims of trafficking assisted by IOM and their accounts of perpetrators. Protected via differential privacy with $\epsilon = 12$.



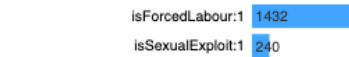
1860 synthetic records matching the query "gender:Female & IP_RecruiterBroker:1 & UN_COE_Region:Europe"

Victim's characteristics

Gender



Type of exploitation



Majority status



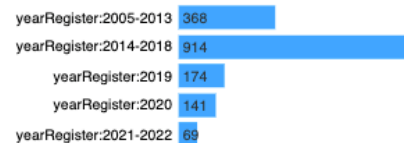
Region of exploitation



Region of citizenship

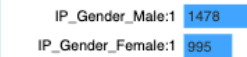


Year of registration

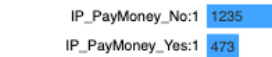


Perpetrator's characteristics

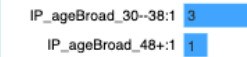
Gender



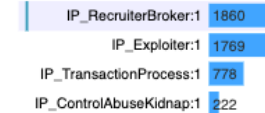
Pay money



Age



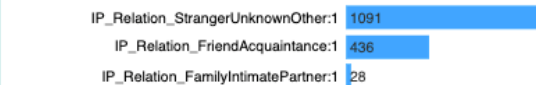
Role



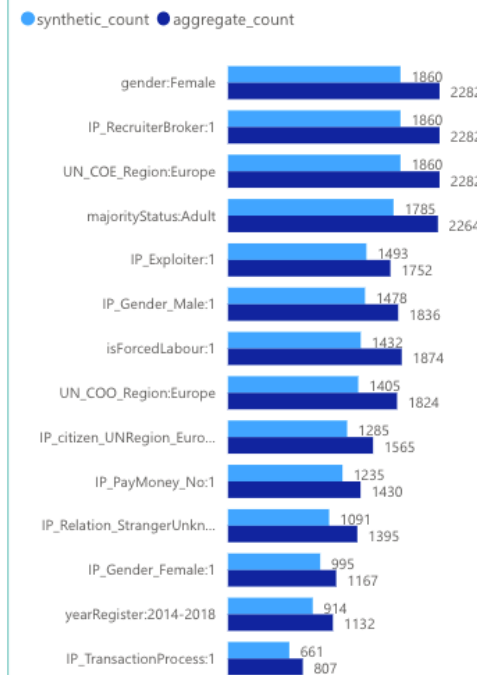
Region of citizenship



Relationship between victim and perpetrator



Comparison of synthetic and aggregate data (based on up to 4 attributes)



Perpetrator's characteristics, e.g.,

- Role
- Region of citizenship
- Victim-perpetrator relationship
- Pay money

Victim's characteristics, e.g.,

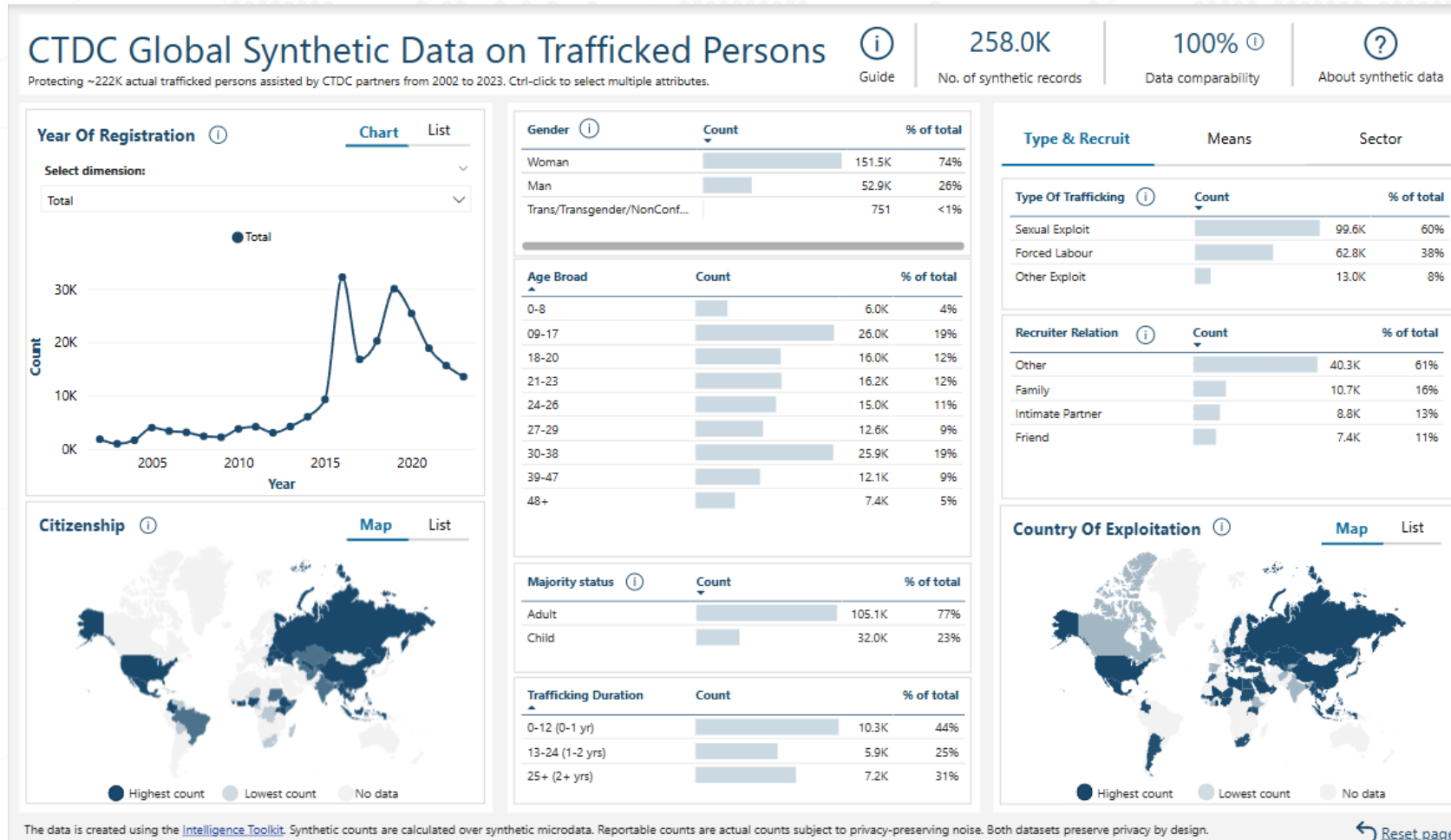
- Type of exploitation
- Region of exploitation and citizenship

Created using [synthetic data showcase](#). Synthetic counts are calculated over synthetic microdata. Aggregate counts are precomputed for short combinations of attributes. Both datasets preserve privacy by design.

Global Synthetic Dataset (2025)

You can download data, codebook, and data dictionary

Protecting over 222K actual trafficked persons assisted by CTDC partners from 2002 to 2023 via differential privacy.




Synthetic Data

Relevance to stakeholders

- More data (that are published safely) can enable more effective research and scalable responses.
- Victim-perpetrator relationships and victims' characteristics can help advance the understanding of risk factors for vulnerability.
- The technology can be used by any stakeholder who wants to collect and publish sensitive data while protecting individual privacy.
- Synthetic data, if well used, can strengthen the evidence base on human trafficking and help address this grave human rights violation.

Want to know more?

- Take this free and self-paced e-learning course on ["Standardized Human Trafficking Survivor Data Management"](#)
- Visit the [CTDC](#) website
-  Read/consult this joint IOM-UNODC report (2023), "Leveraging Administrative Data to Strengthen the Response to TIP (ICSTIP)" and guidance

Appendix: Different forms of data

Type	Definition
Raw data	Data collected on <i>data subjects</i> and not processed .
Partially de-identified data	Data modified only marginally by removing direct identifiers (e.g., name, ID number, IP address).
Aggregate data	Data combined and presented in a summarized format in the form of statistics, etc. (e.g., Global Estimates 2022).
K-anonymized data	Data modified by removing direct identifiers, reducing details (e.g., 23 becomes 18–24), and redacted outliers (e.g., $k = 10$).
Synthetic data	Data that are artificially created rather than obtained through direct measurement, but the statistical properties and relationships from the original data are preserved.

Appendix: Synthesizing data with full k-anonymity (Microsoft)

Step 1

row ID	ATTRIBUTES/COLUMNS			
	A	B	C	D
1	a1	b1	c1	d1
2	a2	b2	c1	d1
3	a1	b2	c2	d1
4	a2	b1	c1	d2
5	a2	b2	c3	d1
6	a1	b1	c3	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cumu. freq.	RAND [0,1]
a1	4	0.25	0.25	0.1
b1	3	0.19	0.50	
c1	4	0.25	0.69	
d1	5	0.31	0.94	
total	16			

new1	a1			
------	----	--	--	--

Step 2

ID	ATTRIBUTES/COLUMNS			
	A	B	C	D
1	a1	b1	c1	d1
3	a1	b2	c2	d1
6	a1	b1	c3	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cumu. freq.	RAND [0,1]
b1	2	0.29	0.29	
c1	2	0.29	0.57	0.6
d1	3	0.43	1	

total	7
-------	---

new1	a1		c1	
------	----	--	----	--

Step 3

ID	ATTRIBUTES/COLUMNS			
	A	B	C	D
1	a1	b1	c1	d1
7	a1	b2	c1	d2

value of attribute	freq.	rel. freq.	cumu. freq.	RAND [0,1]
b1	1	0.5	0.5	

d1	1	0.5	1	0.75
----	---	-----	---	------

total	2
-------	---

new1	a1		c1	d1
------	----	--	----	----

Step 4

ID	ATTRIBUTES/COLUMNS			
	A	B	C	D
1	a1	b1	c1	d1

We could not include attribute B

new1	a1	NULL	c1	d1
new2	NULL	b1	c1	d1

* "freq." stands for frequency. "rel." stands for relative. "cumu." stands for cumulative.