MAKING EACH CASE COUNT

Leveraging administrative data on trafficking in persons

2023

IOM UN MIGRATION

UNODC
United Nations Office on Drugs and Crime

Publisher: International Organization for Migration
17 route des Morillons
P.O. Box 17
1211 Geneva 19
Switzerland
Tel.: +41 22 717 9111
Fax: +41 22 798 6150
Email: hq@iom.int
Website: www.iom.int

This publication was issued without formal editing by IOM.

Required citation: International Organization for Migration (IOM) and United Nations Office on Drugs and Crime (UNODC), 2023. *Making each case count: Leveraging administrative data on trafficking in persons*. IOM, Geneva.

ISBN 978-92-9268-696-3 (PDF)

PUB2021/002/R

MAKING EACH CASE COUNT

Leveraging administrative data on trafficking in persons

This initiative was generously funded by the United States Department of State Bureau of Population, Refugees, and Migration (PRM) and the IOM Development Fund. The contents are the responsibility of the authors and do not necessarily reflect the views of PRM or the IOM Development Fund.

IOM UN MIGRATION | IOM DEVELOPMENT FUND

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF BOXES

LIST OF ABBREVIATIONS

CSO	Civil society organization
CTDC	Counter Trafficking Data Collaborative
GDPR	General Data Protection Regulation
ICS-TIP	International Classification Standard for Administrative Data on Trafficking in Persons
ICT	Information and communication technology
ILO	International Labour Organization
IOM	International Organization for Migration
IT	Information technology
NGO	non-governmental organization
NSO	National Statistical Office
OECD	Organisation for Economic Co-operation and Development
United Nations Trafficking in Persons Protocol	Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime
SDG	Sustainable Development Goal
TIP	Trafficking in persons
UN DESA	United Nations Department of Economic and Social Affairs
UNODC	United Nations Office on Drugs and Crime

GLOSSARY OF TERMS USED IN THE MANUAL

Administrative data	Data collected by various counter-trafficking organizations (e.g. law enforcement agencies, courts and CSOs) as part of their operations
Central agencies	Central government agencies, or other organizations with a coordinating role at the national level, using TIP administrative data from multiple data-producing agencies to produce evidence to address the crime (e.g. national rapporteur's offices, ministries, agencies coordinating the national referral mechanism or national statistical offices)
Civil society organizations	All non-market and non-State organizations outside the family in which people organize themselves to pursue shared interests in the public domain, including NGOs
Data governance	The process of setting the roles and rules for how data are to be managed, including the decision-making process
Data management	The logistics, actual management and processing of data within the rules set by data governance frameworks
Data-producing agencies	Counter-trafficking organizations collecting data on trafficking in persons as part of their operations (e.g. law enforcement agencies, courts and CSOs)
Trafficking in persons	"… the recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs; …" (Article 3 of the United Nations Trafficking in Persons Protocol)

ACKNOWLEDGEMENTS

This manual was jointly developed by IOM and UNODC, with the members of the core team – Harry Cook, Kelly Gleason, Claire Galez-Davis, Marianne Lane, Stine Laursen and Lorraine Wong (IOM); Jesper Bay Kruse Samson, Fabrizio Sarrica and Giulia Serio (UNODC) – benefiting from the generous advice and input of many people.

Particular thanks are due to Darren Edge from Microsoft Research, for his substantial contribution to Chapter V and Annex 6. The team is also grateful to Eduardo Zambrano from the Displacement Tracking Matrix Unit at IOM Headquarters, for his contribution to its work on synthetic data.

The team is grateful to the following IOM and UNODC colleagues for their valuable input and feedback on the draft: Linda Cottone, Samantha Donkin, Estefania Guallar Ariño, Mai Hattori, Natália Maciel, Nuno Nunes, Karla Picado, Antonio Polosa, Verena Sattler, Gabriel Schirvar, Irene Schoefberger, Joseph Slowey, Christina Vasala Kokinakki and Aida Zecevic (IOM); Enrico Bisogno, Salome Flores Sierra, Morgane Nicot, Zoi Sakelliadou and David Rausis (UNODC).

The development of this manual was informed by bilateral consultations with government representatives and IOM and UNODC colleagues. The process ended with a remote workshop in May 2021, during which the draft guidance manual was presented and government representatives and academics took the floor to present their own work. The team owes a great debt of gratitude to all the stakeholders who took part in these consultations and in the workshop, including the IOM and UNODC colleagues who coordinated its consultations with their respective governments. The manual would be much less concrete and practical without them. In turn, it hopes that the manual proves useful to them.

In alphabetical order, the team would like to thank the following: Mary Allen (Statistics Canada), Sofía Arce (IOM Regional Office, San José), Kathy AuCoin (Canadian Centre for Justice Statistics, Statistics Canada), Simon Barrett (New Zealand Ministry of Business, Innovation and Employment), Santiago Baruh (Walk Free), Allen Beck (United States Department of Justice), Oscar Jaimes Bello (National Institute of Statistics and Geography, Mexico), Leila Benaddou (Interministerial Mission for the Protection of Women against Violence and Counter-trafficking, France), Phil Bennett (independent consultant), Jacqueline Bhabha (FXB Center for Health and Human Rights, Harvard University), Alexandra Bonnie (IOM Regional Office, San José), Katherine M. Borgen (Office to Monitor and Combat Trafficking in Persons, United States Department of State), Vanessa Bouché (Bouché Associates), Doreen Boyd (Rights Lab, University of Nottingham), Jessie Brunner (Center for Human Rights and International Justice, Stanford University), Gergana Bulanova-Hristova (Federal Criminal Police Office, Germany), Patrick Burland (IOM London), Matteo Busto (IOM Regional Office, Pretoria), Claudia Cappa (UNICEF), Laura Carpier (IOM Kingdom of the Netherlands), Jose Guillermo Castillo Koschnick (National Institute of Statistics and Geography, Mexico), Frantz Celestin (IOM Nigeria), Katherine Chon (Office on Trafficking in Persons, United States Department of Health and Human Services), Asha Clarke (Public Safety Canada), Kate Cooper (Office on Trafficking in Persons, United States Department

of Health and Human Services), Shirley Cuillierrier (Public Safety Canada), Raquel Damiao (Employment and Social Development Canada), Meredith Dank (John Jay College of Criminal Justice, City University of New York), Elizabeth Darlington (IOM Washington, D.C.), Luis Fabiano de Assis (Brazilian Federal Labour Prosecution Office), Michaelle De Cock (ILO), Richard De Souza (United Kingdom Home Office), Ieke de Vries (Institute of Criminal Law and Criminology, Leiden University), Jenniffer Dew (IOM London), Fatimata Dieng (IOM Senegal), Maria Dimitrova (Bulgarian National Commission for Combating Trafficking in Human Beings), Davina Durgana (Walk Free), Darren Edge (Microsoft Research), Eric V. Edmonds (Dartmouth College), Ahmad Fahim (IOM Canada), Tina Faulkner (United States Department of Labor), Lisa Fischer (Federal Criminal Police Office, Germany), Marlene Fischer (Immigration, Refugees and Citizenship Canada), Vanessa Foronda (IOM Mexico), Alison Gardner (Rights Lab, University of Nottingham), Laura Gauer Bermudez (Global Fund to End Modern Slavery), Alexander Gelovski (IOM Bulgaria), Alexis Gerbeaux (French Ministerial Statistical Department for Internal Security), Elizabeth Gerrior (Polaris), Sara Gilmer (United States Department of Justice), Aileen Girouard (Canada Border Service Agency), Nadine Gies (Federal Criminal Police Office, Germany), Clare Gollop (West Midlands Violence Reduction Unit, United Kingdom), James Goulding (Rights Lab, University of Nottingham), Jessica Gourmelen (Interministerial Mission for the Protection of Women against Violence and Counter-trafficking, France), Lorenzo Guarcello (ILO), Claudia Guidi (United States Department of Labor), Nadia Guiliano (Canadian Centre to End Human Trafficking), Guy Grossman (Political Science Department, University of Pennsylvania), Nana-Ama Gyapomaah (IOM Regional Office, Pretoria), Suze Hageman (Dutch National Rapporteur on Trafficking in Human Beings and Sexual Violence against Children), Javier Hernandez (UNODC Mexico), Thi Hoang (*Journal of Illicit Economies and Development*, Global Initiative Against Transnational Organized Crime), Jacinta Hofnie (Southern African Development Community), Cecilia Hyunjung Mo (University of California, Berkeley), Deyana Ilieva (Bulgarian National Commission for Combating Trafficking in Human Beings), Graciela Incer (IOM Regional Office, San José), Pamela Ingeri (Public Safety Canada), Orla Jackson (Freedom Fund), Phineas Jasi (IOM Headquarters), Duncan Jepson (Liberty Shared), Sarah Johnston-Way (Canadian Centre for Justice Statistics, Statistics Canada), Vera J. Kiefer (Office on Trafficking in Persons, United States Department of Health and Human Services), Saskia Kok (IOM Nigeria), Mónica Lara (UNODC Mexico), Denise Lassar (IOM Regional Office, Vienna), Yuki Lo (Freedom Fund), Abigail Long (Office to Monitor and Combat Trafficking in Persons, United States Department of State), Dilana Lopez (IOM Regional Office, San José), Kate Lytvynets (Microsoft Research), Pieter Maas (IOM Kingdom of the Netherlands), Euan Mackay (Freedom Fund), Alem Makonnen (IOM Regional Office, Pretoria), Cécile Malassigné (Interministerial Mission for the Protection of Women against Violence and Counter-trafficking, France), Paola Martinez (IOM Plurinational State of Bolivia), Anjali Mazumder (The Alan Turing Institute), Mark McCarthy (IOM Regional Office, San José), Craig Melson (techUK), Sélomé Migan (IOM Senegal), Roxane Milot (Global Affairs Canada), Kacy Mixon (USAID), Mariyana Mladenova (Bulgarian National Commission for Combating Trafficking in Human Beings), Elisabeth Moiron-Braud (Interministerial Mission for the Protection of Women against Violence and Counter-trafficking, France), Khaila Montgomery (United States Department of Health and Human Services), Godwin Morka (National Agency for Prohibition of Trafficking in Persons, Nigeria), Nikki Moruti (Southern African Development Community), Alphonci Muradza (Southern African Development Community), Amina Muratovic (IOM Plurinational State of Bolivia), Awa Ndour (National Anti-trafficking in Persons Unit, Senegal), Mody Ndiaye (National Anti-trafficking in Persons Unit, Senegal),

Olatunde Olayemi (Economic Community of West African States), Monica Pardo (IOM Plurinational State of Bolivia), Rita Penedo (Observatory on Trafficking in Human Beings, Portuguese Ministry of Home Affairs), Clara Perez Lopez (IOM Senegal), Dobryana Petkova (Bulgarian National Commission for Combating Trafficking in Human Beings), Elizabeth Pfenning (Office on Trafficking in Persons, United States Department of Health and Human Services), Karla Picado (IOM Regional Office, San José), Ana Catalina Picado (IOM Regional Office, San José), Antonio Polosa (IOM Kingdom of the Netherlands), André Portela Fernandes de Souza (São Paulo School of Economics, Getulio Vargas Foundation), Mathilde Poulhes (Ministerial Statistical Service for Internal Security, Ministry of the Interior, France), Gabriela Rodríguez César (IOM Regional Office, San José), Nerimana Rifatbegovic (IOM Bosnia and Herzegovina), Ashley Russell (Office for Victims of Crime, United States Department of Justice), Rachel Sanchez (IOM Washington, D.C.), Mark Schindel (Public Safety Canada), Irene Schoefberger (IOM Global Migration Data Analysis Centre), Katarina Schwarz (Rights Lab, University of Nottingham), Tamara Sepiurka (IOM Argentina), Bernard Silverman (Rights Lab, University of Nottingham), Wonesai Sithole (IOM Regional Office, Pretoria), Jeni Sorensen (Innovations for Poverty Action), Amandine Sourd (Ministerial Statistical Service for Internal Security, Ministry of the Interior, France), Radoslav Stamenkov (IOM Bulgaria), April Stewart (Global Fund to End Modern Slavery), Shannon Stewart (Global Fund to End Modern Slavery), Hanni Stoklosa (HEAL Trafficking), Agar Tamayo (IOM Plurinational State of Bolivia), Rhidian Thomas (United Kingdom Home Office), Sarah Tietze (IOM Germany), Sharyn Titchener (New Zealand Ministry for Children), Alessandro Tudino (Department for Opportunities, Italy), J.J.M. van Dijk (Tilburg Law School, Department of Criminal Law, Tilburg University), Christina Vasala Kokkinaki (IOM Headquarters), Kyle Vincent (independent researcher), Ilse Waindrich (IOM Kingdom of the Netherlands), Emily Wyman (Rights Lab, University of Nottingham), Monica Zaldivar (IOM Mexico), Veronica Zeitlin (Bureau of International Labor Affairs, United States Department of Labor), Sheldon Zhang (School of Criminology and Justice Studies, University of Massachusetts-Lowell), Cathy Zimmerman (London School of Hygiene & Tropical Medicine).

The team is also grateful to the Intergovernmental Consultations on Migration, Asylum and Refugees for supporting consultations with its Member States, and to Leila Ben Ali and Samson Bel-Aube Nougbodohoue (African Union Institute for Statistics) and Kachi Madubuko and Beakal Bekele (IOM Special Liaison Office in Addis Ababa) for allowing the team to present this initiative to African Union Member States.

The team would also like to thank those who contributed to the editing, layout and design of the report, including Susan Mutti and Joseph Rafanan.

Finally, thanks are due to the United States Bureau of Population, Refugees and Migration and the IOM Development Fund, who financed the development of this manual.

CHAPTER I:
PURPOSE AND SCOPE

There is an acute lack of quality evidence and research available to inform the development of national policies and programmes to combat trafficking in persons (TIP). This is largely due to the lack of data available to researchers and policymakers: trafficking in persons[1] is, after all, a complex, clandestine crime designed to go undetected. That being said, TIP data are collected daily by various counter-trafficking organizations (hereinafter referred to as data-producing agencies – law enforcement agencies, courts and CSOs are common examples) as part of their operations. These agencies produce data of the kind referred to as "administrative data" in the rest of this manual. However, there is no uniform framework or standard practice for measuring the crime of trafficking in persons: indicators on its different aspects vary from country to country, or even from organization to organization within countries, presenting obstacles to comparison and analysis. This, together with resource constraints, unclear division of roles and responsibilities, lack of verification procedures and the sensitivity of the data, makes it difficult to use the data to build a national, regional or international picture of trafficking in persons. Without a robust evidence base, governments and other stakeholders struggle to mobilize the evidence and data to inform and reinforce targeted interventions.

Fortunately, the momentum to improve data collection and analysis, so as to learn more about trafficking in persons, is currently strong, in part thanks to international commitments and pledges (see Box 1). Several countries already have well-established systems to collect TIP administrative data from various points internally and have built databases holding an array of indicators pertinent to TIP cases. Others are now beginning to put such systems in place, often while establishing national referral mechanisms primarily intended for assisting victims of trafficking. Nevertheless, if no steps are taken to ensure that administrative data are collected uniformly and consistently, and are managed and protected by robust systems, the potential risks range from simply misrepresenting the trends, patterns and flows of trafficking in persons to endangering survivors.[2] International guidelines that specify best practices are essential in the effort to produce the highest-quality information and establish proper safeguards from harm.

The purpose of this guidance manual is to support the efforts of governments and other stakeholders to improve data collection, management, sharing and use, so that eventually more high-quality data can be leveraged to inform policy and programming. While the manual will be useful for all stakeholders dealing with administrative data, it specifically targets central government agencies or other organizations with a coordinating role at the national level (hereinafter referred to as central agencies) that use TIP administrative data from multiple sources to produce evidence to address trafficking in persons. These can be national rapporteur's offices, ministries, agencies coordinating the national referral mechanism or national statistical offices, among others. The manual outlines useful considerations, describes the pitfalls to avoid, lists best practices and gives concrete examples to help establish (or improve) all data-related processes for national TIP administrative data. Importantly, it also provides direction on how to use the working version[3] of the new International Classification Standard for Administrative Data on Trafficking in Persons most effectively (ICS-TIP). The ICS-TIP, the companion publication to this manual, establishes a new model of classification for key indicators related to TIP administrative data. It was drawn up to be easily used by any country for the purposes of obtaining, maintaining and

1 The terms "human trafficking" and "trafficking in persons" are used interchangeably in this manual.
2 The terms "victim of trafficking" and "trafficking survivor" are also used interchangeably in this manual.
3 At the time of this manual's publication, the working version is to be put forward for final consultation with United Nations Member States before being submitted for review and endorsement by the United Nations Statistical Commission.

protecting the highest quality of actionable data with which to strengthen government responses to trafficking in persons.

Box 1. International commitments and pledges related to Trafficking in Persons administrative data

The United Nations Trafficking in Persons Protocol

In international law, the offence of trafficking in persons is defined in the United Nations Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, also called the United Nations Trafficking in Persons Protocol. The United Nations Trafficking in Persons Protocol was adopted by the United Nations in November 2000 as part of the United Nations Convention against Transnational Organized Crime. Articles 9.2 and 10.1 pertain to research and the exchange of information.

The Sustainable Development Goals

Three SDG targets explicitly reference trafficking in persons:

SDG target 5.2: Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation.

SDG target 8.7: Take immediate and effective measures to eradicate forced labour, end modern slavery and human trafficking and secure the prohibition and elimination of the worst forms of child labour, including recruitment and use of child soldiers, and by 2025 end child labour in all its forms. (This includes **indicator 8.7.1**: Proportion and number of children aged 5–17 years engaged in child labour, by sex and age.)

SDG target 16.2**:** End abuse, exploitation, trafficking and all forms of violence against and torture of children. (This includes **indicator 16.2.2**: Number of victims of human trafficking per 100,000 population, by sex, age and form of exploitation.)

Reporting on SDG targets must be done using a standard format, in order to gauge progress toward the SDGs.

The Global Compact for Safe, Orderly and Regular Migration

Under Objectives 1 and 10 of the Global Compact for Safe, Orderly and Regular Migration, the signatories make the following pledges:

Objective 1: *strengthen the global evidence base on international migration by improving and investing in the collection, analysis and dissemination of accurate, reliable, comparable data, disaggregated by sex, age, migration status and other characteristics relevant in national contexts, while upholding the right to privacy under international human rights law and protecting personal data. We further commit to ensure this data fosters research, guides coherent and evidence-based policymaking and well-informed public discourse, and allows for effective monitoring and evaluation of the implementation of commitments over time.*

Objective 10: *take legislative or other measures to prevent, combat and eradicate trafficking in persons in the context of international migration by strengthening capacities and international cooperation to investigate, prosecute and penalize trafficking in persons, discouraging demand that fosters exploitation leading to trafficking, and ending impunity of trafficking networks. We further commit to enhance the identification and protection of, and assistance to, migrants who have become victims of trafficking, paying particular attention to women and children.*

The Kyoto Declaration

As part of the 14th United Nations Congress on Crime Prevention and Criminal Justice, which took place in March 2021 in Kyoto, Japan, Member States adopted the Kyoto Declaration, whereby they endeavour to "[s]trengthen efforts to prevent, counter and combat trafficking in persons, including by supporting data collection and sharing as appropriate …".

The ICS-TIP and the present guidance manual are the outcome of a joint IOM and UNODC initiative to support the efforts of governments and other stakeholders to collect, manage and use high-quality, comparable primary data to develop the evidence base on trafficking in persons. Both documents build on, and were informed by, a broad desk review of existing documentation, reports and academic articles, including IOM and UNODC resources developed for their respective operations, Member States and other stakeholders. This includes recent best practices of national and regional data-collection and -management processes. The development of this manual was also informed by direct consultations with governments with varying capacities to collect and manage TIP data, and with experts on TIP data and research. Bilateral consultations took place between January 2020 and April 2021, and a final workshop convened all those consulted in May 2021.

The guidance manual proposes no universal, standardized method of organizing and prioritizing the production of administrative data, simply because there is no one-size-fits-all approach to organizing and institutionalizing anti-trafficking responses or to handling the administrative data produced by organizations playing different roles at different levels in those responses. Rather, it provides guidance for navigating this complex environment and supporting a collaborative, inter-agency ecosystem at the national level that enables administrative data assets to be leveraged ethically, safely and securely to inform anti-trafficking action. It is intended for diverse environments with different levels of resources and capacity. Where the text requires technical language to describe data processes, the document provides examples and detailed definitions of concepts, so as to be accessible to users with limited technical backgrounds.

CHAPTER 2:
SUMMARY

WHY ARE ADMINISTRATIVE DATA SO IMPORTANT IN THE FIGHT AGAINST TRAFFICKING IN PERSONS?

There is one big question weighing heavily on the minds of stakeholders: "what works" to end trafficking in persons? The answer is of equally pressing concern to front-line assistance workers, local and national policymakers, and donors at all levels. Governments need evidence to allocate resources and formulate the best policies to reduce the incidence of trafficking locally and abroad, across all four Ps: prevention, protection, prosecution and partnership. For example, CSOs need to know how best to deliver services, facilitate the exit of victims from exploitative conditions and prevent vulnerable individuals from being trafficked. Law enforcement agencies need detailed information on recruitment processes and perpetrator profiles, and precise geographical information about where trafficking is likely to take place. Better knowledge of the crime can also help prosecutors build their cases and inform international, regional and bilateral cooperation. Donors, from the grassroots all the way up to major official development assistance contributors, must know where spending can have the best impact. All of these questions require a more comprehensive, complete evidence base, the building blocks for which are data. Many governments have agreed or committed to collect and share data on trafficking in persons to fulfill national and international commitments (see Box 1 for a summary of the main international commitments).

Fortunately, as mentioned earlier, data are collected by various organizations as part of their day-to-day operations. In fact, the growing demand for data has catalysed an exponential increase in initiatives to produce new data sources and/or digitize/consolidate existing case records.

OBSTACLES TO THE PRODUCTION OF HIGH-QUALITY ADMINISTRATIVE DATA

Nevertheless, there remain obstacles to the collection, use and sharing of centralized high-quality data. Some of these obstacles are described below.

1. *The absence of standardized indicators*

   There is no uniform framework or standard practice for measuring trafficking in persons. Indicators on its different aspects vary from country to country, or even organization to organization within countries, hampering comparison and analysis. More broadly, without clear, standardized definitions of concepts, there is no way of knowing whether the data collected from various sources (NGOs, different jurisdictions, different countries) or over different time periods actually pertain to the same phenomenon.

2. *Data privacy concerns*

   When data are collected from consenting individuals, the individuals must be protected from the risk of harm they might be exposed to if the data are shared. When the individuals are members of a vulnerable or exploited group, the risk of harm is often much greater than for other data subjects. Without robust rules and systems of accountability for protecting this information, data collection, storage and sharing pose risks for all involved.

3. *Coordination challenges and capacity constraints*

   Even when the data do exist, they are often housed in siloes between different CSOs, government offices or jurisdictions. Coordinating their use is made difficult by various factors, including the aforementioned data privacy concerns. Unclear divisions of roles and responsibilities between relevant authorities also create coordination challenges. Resource constraints and a lack of tools and guidance hamper the ability of countries to invest in information technology, staff training or statistical or data management policies and procedures.

4. *Data quality*

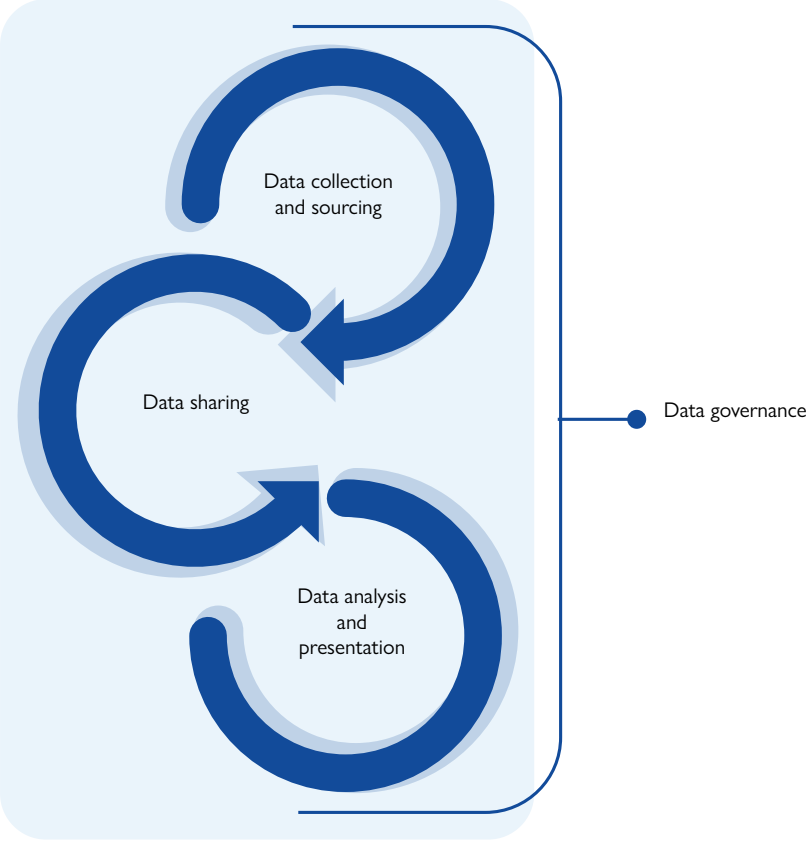
   The quality of the data collected is often impacted by faulty procedures, lack of verification standards or failure to triangulate findings. Overlapping coverage of data sets or imprecise criteria for sourcing information may result in less accurate data, and in the worst cases, misleading statistics, misrepresenting national trends, patterns and flows. Finally, even if all steps are taken to ensure adequate data collection and management procedures are followed, the data may be miscommunicated in a way that is counterproductive to the intended aim (for example, causing misrepresentation of trafficking occurrence and/or misallocation of resources).

The above list, while non-exhaustive, covers common issues that can impede the generation and use of quality data in general, including for policymaking. Some characteristics particular to TIP data, however, can further compound these problems: for example, the risk that victims will be revictimized if data privacy measures are not properly considered. In addition to this, the response to trafficking in persons tends to involve many different stakeholders, within and across countries, which can complicate coordination efforts.

SUMMARY OF CONTENTS

All of the challenges in the collection and management of quality data listed above occur at one or more stages in the data life cycle. While variations exist, in the case of TIP administrative data, this cycle essentially consists of the initial collection or sourcing of data, data sharing or publishing, and data analysis and the subsequent process of presenting and interpreting the results obtained from the data (see Figure 1). The management of data throughout these stages requires a data governance framework.

Figure 1. The data life cycle



Source: Unless otherwise specified, all figures and tables have been produced by IOM and UNODC.

While some overarching principles of data protection and harmonization are common to all data governance guidance, other components are highly specific to the type of data being referenced. The challenges of data collection, management and use particular to TIP data, detailed above, require specific consideration.

Chapters III to VI of this manual provide practical guidance, broadly following each element of the data life cycle.

Chapter III: Data collection

The first step in the data life cycle – data collection – lays the foundation for all subsequent steps. Unless well-planned standard operating procedures, embedded ethical norms aligned with relevant legislation and policy frameworks are present from the outset, it is not possible to produce quality administrative data that can be used to generate standardized,

comparable and consistent indicators of trafficking in persons. All the efforts to analyse, protect or communicate the data cannot offset the initial collection of poor-quality data.

One of the foundations for planning data collection is to be clear about what data are needed and for what purpose. Chapter III starts with some examples. It then identifies common issues facing the central agencies tasked with sourcing administrative data and explains how to obtain better data. This part of the chapter describes what good data should look like and how the ICS-TIP can be used to get there. Applying the ICS-TIP will not suffice, however, to ensure quality data: how the data are collected is in many respects even more important. Chapter III therefore details the ethical principles of data collection, including the "do no harm" principle. It also considers planning which organizations to work with to collect TIP administrative data at the national level, how to establish good partnerships and coordination to encourage organizations to share data, and issues around bringing very different types of data together. It ends with considerations on improving the capacity of data-producing agencies.

Chapter IV: Data governance

Data are managed according to specific rules, from collection to sharing and use.[4] These rules are set within a data governance framework,[5] so as to encourage the production, sharing and consolidation of administrative data from multiple agencies at the national level.

The data governance framework should stipulate policies and procedures to safeguard data assets/subjects (or harmonize and coordinate them, if they already exist at the institutional level) and mitigate risks and challenges related to privacy concerns, data organization and the harmonization of TIP administrative data between sources. The advice provided in Chapter IV is more practical than technical: the focus is on useful considerations and questions that the roles and rules stipulated in a data governance framework need to address, rather than on specific types of software or a specific framework to be applied in all contexts.

Chapter IV starts by listing some of the main objectives of a data governance framework and then outlines the roles and the accompanying responsibilities that need to be created as part as that framework, together with the rules that need to be established. Who decides who can access the data, and according to which rule? Who, or what, decides whom to share data with, and which data? While there is no single answer to these questions, the data governance framework needs to be able to answer them. Chapter IV then considers how these issues can be applied in the kind of inter-agency environment commonly encountered when compiling TIP administrative data at the national level.

Chapter V: Sharing and de-identifying administrative data

Chapter V starts with some general, theoretical considerations on data protection when it comes to sharing and publishing data; specifically, the issues involved in de-identifying sensitive – and particularly personal – data. In most cases, there is a trade-off between data utility (i.e. how much can be learnt from the data) and privacy (i.e. how detailed the data are, and/or how easy it may be to identify someone's data). While much of the data reported at

[4] Data governance is the process of establishing the roles and rules for how data are to be managed, including the decision-making process. Data management refers to the logistics and the actual management/processing of data within the rules set by data governance frameworks.

[5] That is to say, the rules and roles that govern the management of data.

the national level is presented in aggregate form,[6] it is the granularity of disaggregate data that is most useful for the purposes of analysis. Different data are needed for different purposes by different types of user; the utility/privacy trade-off thus needs to be evaluated for each type of data and each type of user. Chapter V defines what is meant by personal data (including direct and indirect identifiers) and de-identified data (including different "levels" of de-identification). The remainder of the chapter is devoted to good practices for publishing and sharing data. The penultimate section describes different methods of de-identifying data and some of their advantages and drawbacks. The last section highlights some due-diligence considerations that apply every time data are shared and published.

Chapter VI: Administrative data analysis and presentation

Communicating data responsibly is the last stage in ensuring that quality data accurately inform policy and programming. Even consistent, reliable data are useless if the results derived from them are not understood and interpreted correctly, and hence misinform the counter-trafficking response or other stakeholders. Chapter VI provides specific examples of common mistakes. To avoid them, it is crucial to understand what kinds of conclusion can be drawn from the analysis of administrative data.

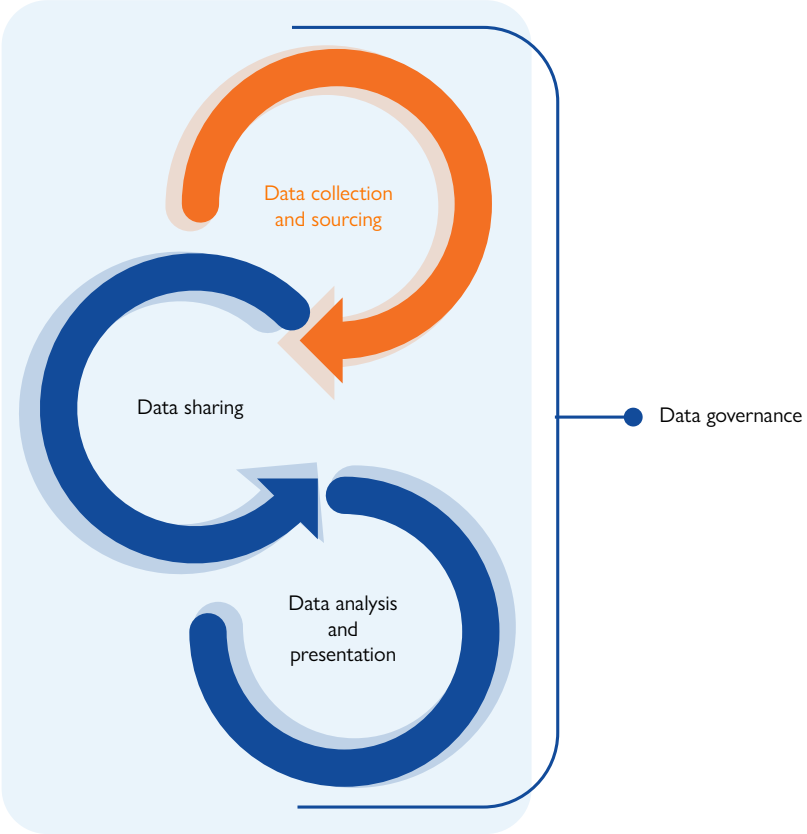
With that in mind, the first half of the chapter is devoted to understanding the place that administrative data occupy within the wider evidence landscape and other TIP sources of data. Chapter VI also describes in detail the strengths and limitations of TIP administrative data, and contrasts these with those of other data sources, such as surveys, geospatial data, big data and qualitative data. The second half of the chapter describes the good practices to apply and the pitfalls to avoid when presenting TIP administrative data.

[6] Which can give rise to issues of its own, as will be seen in Chapter V.

CHAPTER 3:
DATA COLLECTION

Data creation, or the collection, sourcing or capture of data, is the first stage in the data life cycle (see Figure 2). Data quality is highly dependent on this initial step in the process. Problems of data inaccuracy or incompatibility cannot be corrected during data storage, management and sharing unless each relevant indicator has all the requisite properties of high-quality data collected correctly.[7]

Figure 2. The data life cycle: data collection



Central agencies, which are responsible for compiling TIP administrative data from multiple sources, are rarely on the front lines of the initial data-collection process. Whether the administrative data are needed to discharge their mandate to combat trafficking, for reporting obligations or to establish a national repository, officials may need to source them from a range of external partners, including front-line assistance organizations, police forces, shelters and hotlines, or from intergovernmental agencies.

In this scenario, even though most central government agencies obtain TIP data second-hand, the primary data-collection process remains a matter of concern to them. If the initial data are collected improperly, unethically or outside the legal framework, they may be unusable at best and harmful at worst.

This aim of this chapter is to provide guidance on effective sourcing of TIP administrative data assets by central agencies. The content below describes the processes involved in the thorough and ethical collection of high-quality data. The overarching themes include how specific data types can fulfill government objectives, barriers to data collection and the ways in which governments can offer support to facilitate better data collection. The chapter

[7] See J. Brunner, *Getting to Good Human Trafficking Data: Everyday Guidelines for Frontline Practitioners in Southeast Asia* (Stanford University, 2018), pp. 11–12.

synthesizes existing guidance, especially concerning ethical data-collection standards, and presents principles, challenges and best practices.

Importantly, this chapter introduces and illustrates the new ICS-TIP, the companion document to this guidance manual. It draws particular attention to the establishment of standardized indicator definitions. It covers the range of individual, event-related and organization indicators needed for various stakeholder purposes and identified in the ICS-TIP, and discusses how governments can build capacity for more advanced levels of data collection in all contexts.

Finally, this chapter contains practical recommendations for how central agencies can support a collaborative, inter-agency data ecosystem that better serves all stakeholders. The last section covers the development of protocols and systems to enhance coordination among various national data stakeholders, with a view to streamlining the process of harmonizing administrative data for its use for evidence purposes.

THE NEED FOR ADMINISTRATIVE DATA ON TRAFFICKING IN PERSONS

Data on trafficking in persons cases and the individuals involved are highly sensitive and can be difficult to obtain. Governments have to collect data on the many dimensions of the issue with the help of multiple data-providing stakeholders.

In order to minimize the risk to data subjects and not place undue strains on CSOs and government resources, it is key to take a targeted approach to collecting the specific data needed to meet core government objectives.

Gathering the right types of data to achieve government objectives

To be useful, data have to be relevant to the objectives.[8] This means that, from the very start of the data-sourcing process, government agencies must take stock of their needs in order to determine which data will best serve their purposes.

The nature of the crime of trafficking in persons is such that many types and sources of data will likely be needed to develop a robust national (or even local) picture of the phenomenon. Tackling the crime involves gaining insights on the perpetrators as well as the victims, gathering information on acts, exploitation, relationships and locations, discovering the mechanisms used for recruitment, methods of coercion, common industries of exploitation and so on.

TIP data collection at the government level has a host of primary objectives (see Table 1). Most people who are familiar with the TIP data landscape can attest to the fact that obtaining any one of these data types, let alone all, would require substantial resources. In many contexts, several of these data assets are unlikely to exist.

Administrative data are primarily produced as a by-product of various operational processes or to support their delivery (e.g. to serve as a record of services, assessments or decisions; document legal entitlements or protections; provide a trail of accountability; or show how public funding is allocated). These rich sources of data have the potential to serve many of the government objectives listed in Table 1. Assisting government agencies to leverage these data assets so as to inform anti-trafficking responses, while challenging, is what this manual sets out to do.

[8] UNODC, *Toolkit to Combat Trafficking in Persons* (Vienna, 2008), p. 474.

Table 1. Data types needed to meet specific government objectives

Data type	Objective
Data on trafficking in persons events/processes are needed …	… to improve knowledge of the scale and extent of trafficking, and to formulate an adequate response.
Demographic information on victims and perpetrators is needed …	… to understand the nature of trafficking in human beings, identify new trends and vulnerabilities, and target projects and programmes at those in need.
Demographic information on groups in vulnerable situations is needed …	… to develop (prevention) projects and programmes, and to reduce risks.
Information on the means, act and purpose of the trafficking event is needed …	… to understand trends in the crime of trafficking in persons (including confirmation/identification that the event is a case of trafficking), recommend actions, and target projects, programmes and law enforcement efforts.
Nationally representative prevalence estimates are needed …	… for international reporting and to improve knowledge of the scale and extent of trafficking.[a]
Organizational information on programming (impact) (monitoring and evaluation) is needed …	… to develop projects and programmes, enhance the relevance of training programmes and help reduce risks.
Policy (impact) data are needed …	… to recommend actions, and to target projects, programmes and law enforcement efforts.[b]

Source: IOM, *Guidelines for the Collection of Data on Trafficking in Human Beings, Including Comparable Indicators* (Vienna, 2009). For the purposes of this manual, the original list has been condensed into core objective areas.

[a] Administrative data can provide much insight into patterns and characteristics of trafficking events and help build an understanding of those victimized by and perpetrators of the crime, but they cannot paint a complete picture of prevalence. However, if the right data are collected from multiple national sources, new methods of estimation can utilize case data to produce nationally representative estimates.

[b] Policy data, especially those that assess the implementation and/or effectiveness of policies, are extremely valuable, but often difficult to obtain. While data on legislation and policy are easy to come by, data on policy implementation are harder to obtain, especially at local level. Connecting policy implementation to lower crime rates and/or reduced risk/vulnerability would require time-series data on prevalence before and after the policy was introduced, or good comparable data on prevalence between locations that vary in terms of policy action. Such data are a more distant prospect than other data types.

COMMON CHALLENGES FACING DATA-SOURCING AGENCIES

Central agencies encounter numerous challenges when they compile administrative data from multiple data-producing agencies, as a consequence of both the nature of first-hand administrative data collection and the specific challenges related to the crime of trafficking in persons. Consultations with government experts on data sourcing revealed that the areas discussed below were among the main stumbling blocks.

Capacity of data-producing agencies

The clearest challenge lies in the fact that the first-hand collection of this type of data rests in the hands of the various agencies and organizations that produce administrative data, such as front-line protection agencies, which can face a variety of data-collection challenges.

Unlike administrative data-producing agencies, academic researchers collecting primary data do so as part of carefully designed research plans, which usually need to be submitted to and approved by an independent review board, to ensure that the collection strategy is ethical and sound. When it comes to the collection of TIP administrative data, the process on the ground is rarely as straightforward. The challenges to comprehensive data collection (that also adheres to ethical standards for the protection of data subjects) are numerous.

First, these include setting up data management procedures, tools and systems (even for paper records) and, in the long run, ensuring those systems are maintained and supported. These resources, and the knowledge required to manage them, are often not available to front-line data collectors. For example, some organizations may be less likely to process data on children generally, because children cannot consent to data processing and the organization may not have the immediate capacity to conduct a best-interest determination and may still have to clarify issues of guardianship and safeguarding at the time of contact.

In addition, even if workers in data-producing agencies have some training in data collection, another pressing obstacle remains: It is a secondary objective beyond the core responsibilities of a role that is often difficult and can be dangerous. Any data collection expected of front-line responders cannot be so involved or time-consuming that it interferes with their primary role. It is important to understand the time constraints of these workers when prioritizing the types of data needed to meet government objectives. This is particularly true in emergency contexts.[9]

Relatedly, while workers in data-producing agencies are highly experienced in serving individuals and local communities, and have in-depth knowledge of the issues facing them, with a background in and experience of rigorous data-collection methods. They may not be able to consistently apply strict definitions to, or even record, all data fields for the cases of trafficking in persons they document, particularly if those data fields are primarily requested for research rather than protection purposes.

Heterogeneous data sources

Central agencies receive highly variable data sourced from multiple organizations, agencies and institutions. Four main reasons explain this variability.

The first relates to **how organizations interpret the law to identify a case of trafficking in persons**. Even though they all operate under the same legal definition of trafficking in persons, different organizations may operationalize this definition in different ways. UNODC assessments of practitioner attitudes show that the practical interpretation of certain aspects of the definition differs according to jurisprudence, cultural context, policies applied and so on. For example, what counts as abuse of a position of vulnerability may be subject to interpretation, as does the extent to which potential victims have to demonstrate

[9] IOM has published guidance on how to integrate counter-trafficking data collection and analysis into existing information management mechanisms in such contexts: IOM, *Counter-trafficking in Emergencies: Information Management Guide* (Geneva, 2020).

that they have been under the control of someone else, if it is clear that they have been exploited. Additionally, the law is not intended to serve as a guide to data collection. Legal definitions are designed to be broad enough to capture a range of different criminal activities and trafficking in persons can manifest itself in many different ways; the law is not designed to specify what data should be used to describe those different manifestations of the crime from a research perspective.

The second main reason for the heterogeneity of data is related to **how organizations record information once a case is identified**. Different organizations have different roles to play. Some organizations may have a criminal justice mandate, others a protection mandate, others yet a law enforcement mandate, and this will affect what data they record and on whom. For example, some organizations will record means of control for children, others will not, or not consistently. This is because such means need not be demonstrated to identify victims when they are children. In another example, some organizations will collect information on perpetrators, others will not.

The third reason for heterogeneity is that **organizations differ in the formality of their assessment or identification determinations**. In many countries, only specific organizations are mandated formally to identify victims of trafficking such that they can avail themselves of the protections offered by the State (e.g. the right to remain in the country for a period of time and immunity from prosecution for acts undertaken while being trafficked), even though many front-line agencies may identify and assist victims and potential victims regardless of whether they have been formally recognized as such by the agency. Accordingly, a central agency may collect data from "formal" sources and/or "informal" sources, raising the question of how, and whether, to prioritize and/or harmonize these data. Are more "formal" sources likely to be more trustworthy than "informal" sources? This is not necessarily the case, as many organizations are highly skilled at identifying victims of trafficking, even if they lack a more formal mandate to recognize the status of an individual or case.

Relatedly, **organizations differ in their likely coverage of the (un)identified population**. Some types of organization may be more likely to identify, or come into contact with, some cases of trafficking than others and this is often linked to their mandate and operations. The threshold of proof for conviction in a court is typically higher than that asked of victims to demonstrate that they are eligible for CSO protection and support services. Consequently, cases are only brought to court where sufficient evidence exists and where there is a good chance of successful prosecution. This means that the number of cases prosecuted in court tends to be a smaller subset of the total cases identified by all anti-trafficking entities in a given context. The victims may also be too afraid to approach the police or pursue prosecution because they are scared about possible reprisals from perpetrators or criminal networks. They may also be in irregular migration situations and simply worried that they may themselves be subject to action by border or law enforcement agencies. These dynamics can mean that CSO protection agencies are more likely to encounter some cases of trafficking than the police and that the CSO sector as a whole may identify a larger subset of the population of trafficked persons than the police, for example. Finally, an agency may be more likely to identify certain types of trafficking if it has a specific mandate or is specialized in addressing specific forms of exploitation, such as labour exploitation in the case of a labour inspectorate.

Figure 3 provides examples of where different types of organizations might fall on the coverage/formalization quadrant. For instance, court and prosecution data will usually have a high level of formality, but may only capture few individuals. On the other hand, CSOs specialized in assistance will likely have greater coverage, against low formality.

Figure 3. Data sources: general considerations on formalized identification versus coverage

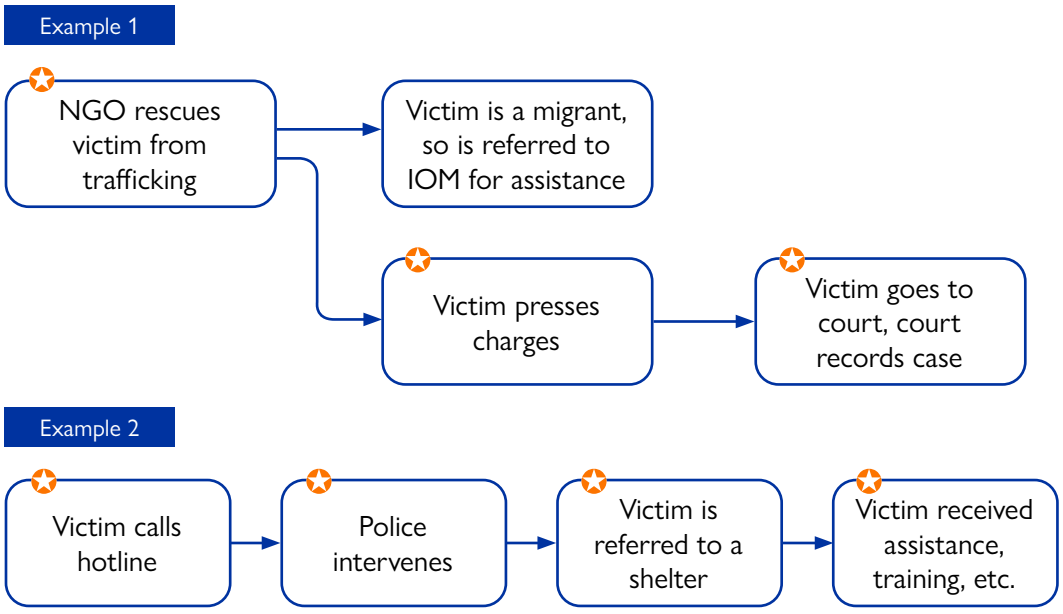
		Coverage of cases		
		Low	Medium	High
Formalized identification	Low	✓ Non-specialized front-line agencies	✓ Specialized CSOs	
	Medium		✓ Labour inspectorate	✓ National referral mechanism
	High	✓ Court/ prosecution	✓ Specialized police units	

Overlapping coverage of data sets

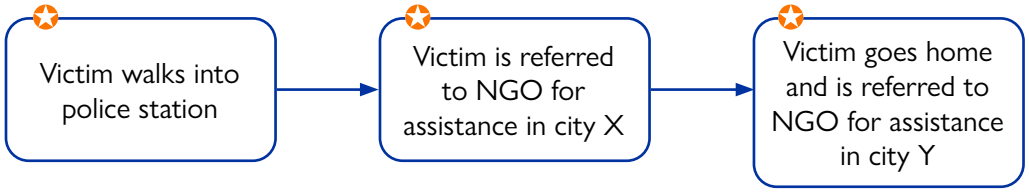
Overlapping coverage of data sets is perhaps the most common dilemma faced by data-sourcing agencies. The previous section highlighted how organizations differ in their coverage of the TIP population, although it is likely that there will be some degree of overlap. However, when data assets are obtained from multiple sources, it is often unclear to what extent there is overlapping coverage of trafficking cases between data sources. For example, the same individual victim of trafficking may encounter multiple front-line agencies in their search for appropriate services and counselling. Each point of contact can be recorded in different ways by front-line agencies, which may obtain different information with potentially no common identifiers to link the records or even determine that they concern the same person.

Figure 4 provides several examples of different points of contact that a victim may have with different front-line agencies, where administrative data may be recorded and processed.

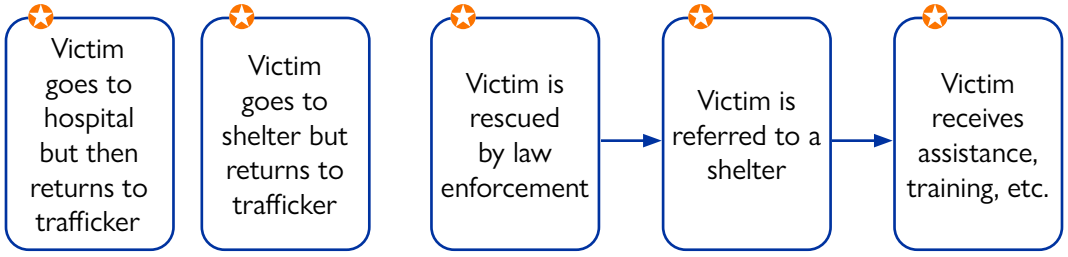
Figure 4. Multiple processes and sites of contact/data-collection points*



Example 3



Example 4



* The star indicates each point at which data on victims and aspects of the crime can potentially be collected. Note that these examples are hypothetical and simplified to illustrate the possible administrative stages where data might be recorded or processed. In reality, case management is often more complex and should be approached with the specific case and context in mind.

As can be seen in examples 1, 2 and 3, without effective coordination on data collection, all the chains of assistance clearly have the potential to create new administrative records on the same individual or trafficking event, which could lead to it being counted multiple times. Example 4 is of special interest, as it shows a victim falling in and out of the care system, re-emerging at different points in time in what may be a case of re-victimization; this is qualitatively different from multiple agencies re-counting the same event/case.

Ensuring the legal basis for using data downstream is planned for in upstream primary data collection

Central agencies wishing to use administrative data produced by other organizations' operations are essentially third parties who sit "downstream" in the data life cycle and wish to use data produced by others "upstream" for their own purposes. In this case, the downstream purpose for which the third party wants to use the data is to produce evidence to inform anti-trafficking activities. This may not be the primary purpose for which the administrative data are originally produced upstream, which would be to support the day-to-day operations of the data-producing organization. This could be supporting the delivery of protection services for victims, in the case of a front-line CSO, for example.

It is important that all purposes for which the data are intended to be used downstream are planned for and supported by the collection and processing approach upstream. That is, the legal basis upon which data are collected and processed upstream must encompass downstream usage by a third party (the central agency) for a specific purpose (producing evidence to address trafficking in persons). This includes ensuring that any rights that an individual or organization has over the data are upheld in the process. These considerations are all the more acute in the case of personal data, which are highly sensitive, usually protected by specific legislation, and where it is essential that the rights of data subjects are upheld.

The possible legal bases for collecting and processing personal data will depend on the national laws in place, but examination of the United Kingdom's General Data Protection

Regulation, which is based on the European Union GDPR, provides an instructive example.[10] Processing of personal data in the United Kingdom must be based on at least one of the Article 6 categories[11] (multiple legal bases may be possible). These include, but are not limited to:

a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;

…

d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;

e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

The full list of Article 6 categories is available in Annex 1. The first basis listed, "consent", is described in greater detail below but can briefly be said to involve informing data subjects about the specific purposes for which their data are being collected and letting them choose whether their data can be collected and used for those purposes. One key point is that if the legal basis for original, upstream, collection or processing is consent, then all downstream processing must also be on the basis of consent. The second basis listed, "vital interests", refers to situations in which personal data need to be processed in order to protect someone's life and there is no other way to protect them that would be less intrusive. The "vital interests" basis cannot be used "for health data or other special category data if the individual is capable of giving consent, even if they refuse their consent". The third basis listed, "public interest", is relevant to public authorities and organizations with "official authority" (covering "public functions and powers that are set out in law"), but also to any organization carrying out tasks in the public interest.[12] There must be no other way to undertake these tasks that would be less intrusive. Regardless of which legal basis is relevant, it is important to document the decision(s) that is(are) taken in this respect, for accountability to all stakeholders and to document compliance, should it be requested.[13]

While administrative data on trafficking in persons are often (but not always) personal data or derived from personal data, some may not be just "any" personal data, particularly when it comes to TIP victims. The United Kingdom GDPR also contains special provisions for categories of personal data that are particularly sensitive. Two categories of personal data that are deemed highly sensitive and for which special protections apply are special category data and criminal offence data. The former category refers to personal data needing more protection because of their sensitivity (e.g. racial or health data), while the latter refers to "personal data relating to criminal convictions and offences or related security measures".[14] The legal bases set out in Article 6 will not suffice to process these data. For special category data, one of the specific conditions in Article 9 must also be met. These conditions include "[r]easons of substantial public interest (with a basis in law)", "[h]ealth or social care (with a basis in law)" and "[a]rchiving, research and statistics (with a basis in law)", which are often relevant for TIP data (with reasons of substantial public interest including, for example,

10 Further instructive examples are to be found in Australia's Privacy Act and Mexico's Data Protection Law.

11 See www.legislation.gov.uk/eur/2016/679/article/6.

12 Specifically: "You do not need a specific statutory power to process personal data, but your underlying task, function or power must have a clear basis in law." See the United Kingdom Information Commissioner's Office description of "public task".

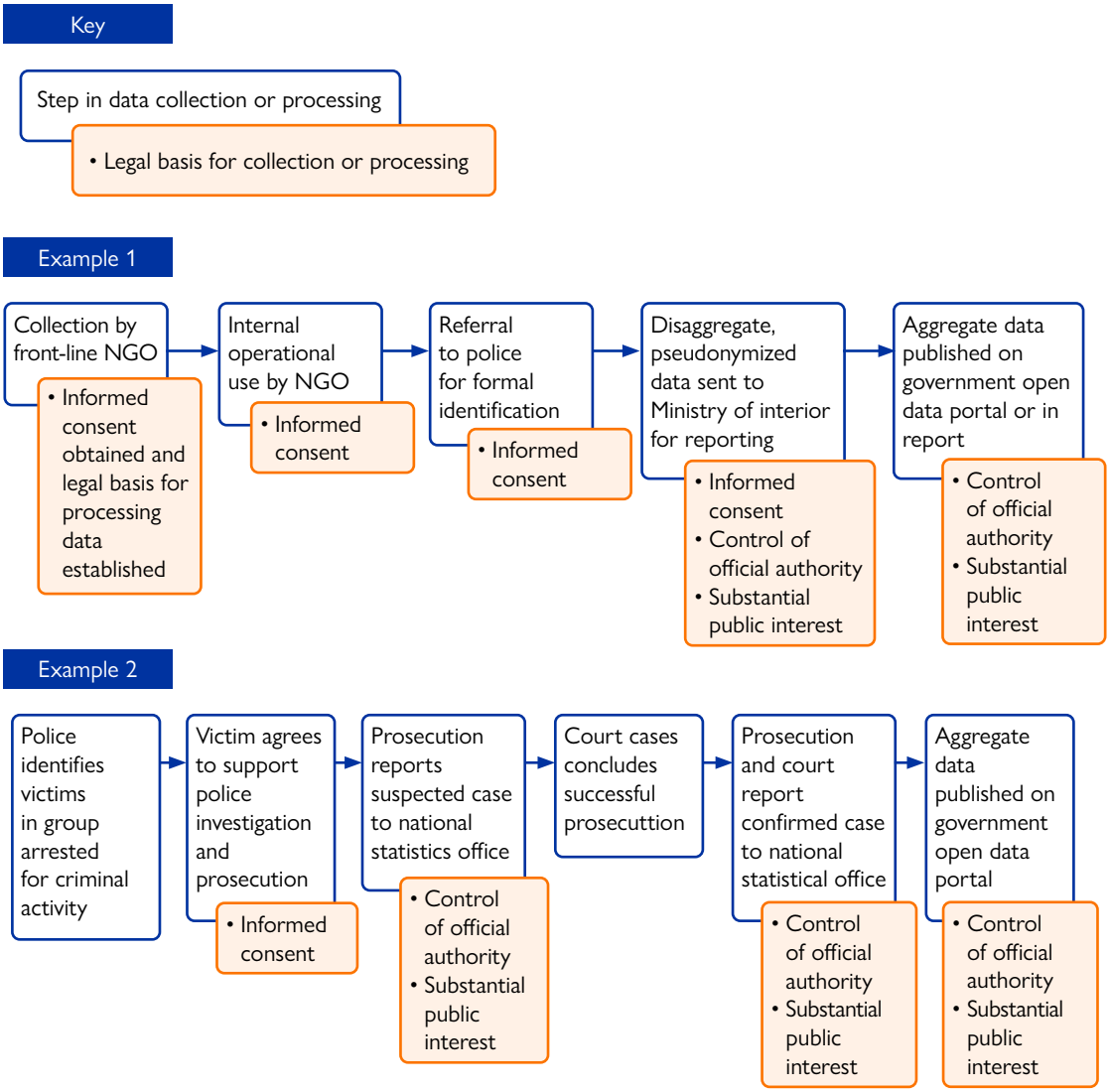
13 More information on the lawful basis for processing is provided on the website of the United Kingdom Information Commissioner's Office, on which this paragraph draws extensively.

14 See the website of the United Kingdom Information Commissioner's Office for more information on special category data and criminal offence data.

safeguarding children and individuals at risk and preventing or detecting unlawful acts).[15] For criminal offence data, it is necessary to have official authority or meet one of the conditions set out in Schedule 1 of the Data Protection Act (these conditions include research and preventing or detecting unlawful acts, for example).[16] More information on both special category and criminal offence data is available in Annex 2. Administrative data assets that are useful for producing evidence on trafficking in persons may include both types of data.

Drawing from the example of the United Kingdom, Annex 3 provides two examples of data flows and the legal basis enabling each step in the pipeline, also summarized in Figure 5 below. The first example shows that, when consent is the legal basis for collecting the data, it may not need to apply in subsequent steps, particularly once de-identified derivatives have been produced. On the other hand, the second example relies on control of authority (and substantial public interest).

Figure 5. Two examples of data pipelines with each step's legal basis



[15] See the website of the United Kingdom Information Commissioner's Office for more information on Article 9) and on the "substantial public interest" conditions.

[16] For more information on Schedule 1, see the website of the United Kingdom Information Commissioner's Office at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/criminal-offence-data/#schedule1 and www.legislation.gov.uk/ukpga/2018/12/schedule/1/enacted.

Robust legal frameworks are essential for protecting data and upholding the rights of data subjects. They should be viewed as firm foundations rather than barriers to data collection and processing. Nevertheless, as the above examples illustrate, central agencies aiming to handle a range of sensitive TIP administrative data assets from various organizations will need to work with those organizations to plan each data pipeline rigorously. In many cases, it is also beneficial for central agencies to have their purpose clearly established or enshrined in law. This is the case of Statistics Canada, for instance, as explained in the section "How to establish trust with data-producing agencies".

Differences in legal definitions

Another challenge to overcome in the effort to source good, standardized data relates to different definitions of trafficking in persons, even within a single country.

Consider, for example, the data collected on the incidence of trafficking in persons in the United Kingdom. The United Kingdom was one of the first countries to adopt data-collection protocols and systems to support its national referral mechanism. Nevertheless, according to Home Office reporting of the quarterly data on modern slavery, data collected in England, Wales, Scotland and Northern Ireland cannot be combined.[17] Expert consultations reveal that the same issue appears to be a major problem for federal governments, including in the United States of America, as individual States may have significantly different legal frameworks and there may be coordination challenges between them.

The issue is the same, to some extent, for regional and international comparisons. According to UNODC, national legal definitions of trafficking in persons are broadly in line with the international definition set out in the United Nations Trafficking in Persons Protocol.[18] Nevertheless, "broadly in line" leaves room for interpretation, as explained by UNODC:

> The potential breadth and narrowness of the definition has raised several issues that States have taken quite different positions on. There is a tension between those who support a conservative or even restrictive interpretation of the concept of trafficking, and those who advocate for its expansion.[19]

Of course, this complicates efforts towards regional (e.g. European Union, regional economic communities) and international harmonization and compilation.

FACILITATING THE SOURCING OF ADMINISTRATIVE DATA

In reality, for all the reasons discussed above, the fact that high-quality data are difficult to collect creates problems for government agencies sourcing data. However, with increased support from and coordination with such agencies, many of these problems can be reduced, better managed and eventually overcome.

Properties of better-quality administrative data

Administrative data need to be of a certain quality and have particular properties if they are to be of optimal use for government evidence purposes. Due to the scarcity of data on trafficking in persons, it is important to maximize the value of all potential data assets. This

[17] Office for National Statistics, Modern slavery in the UK: March 2020.
[18] UNODC, *Global Report on Trafficking in Persons 2020* (United Nations, New York, 2020), p. 61.
[19] UNODC, *The International Legal Definition of Trafficking in Persons: Consolidation of research findings and reflection on issues raised*, Issue Paper (United Nations, Vienna, 2018), p. 2

includes ensuring that data are reliable and that the diverse data sources can be combined and used together in a meaningful way.

Accuracy and standardization

Data quality has many components but, overall, high-quality data can be trusted as a reliable, accurate resource. One major factor of reliability is knowing that the source of the data is authentic (unaltered and unbiased). Another is that the data are precise, or recorded properly and uniformly.

A related characteristic of quality is standardization. Standardized data have a common definition, meaning that the way something is measured does not change no matter who is collecting the data or where it is being collected. Simply put, when all data-producing agencies follow the same protocols, recording TIP indicators in exactly the same way, no matter which part of the country they reside in and which department or CSO is doing the collection, the data are standardized.

Establishing high-quality, standardized indicators is the first step in building understanding of trends in and characteristics of trafficking in persons. To evaluate trends, governments need consistent data collection over time. Consistency requires that the same standard-based indicator measurement classifications be used and that data are collected for the same set of indicators at regular intervals.[20]

Interoperability

The related concept of interoperability implies that the technical aspects of different data sources allow them to be easily combined. Data must be standardized, or at least compatible, to be interoperable, but they must also have unique identifying information, so that cases are not counted more than once and different data sources can be used in combination with one another. In addition, by using unique identifiers to link data sets, a more complete picture of the TIP situation will emerge over time, including cases of re-victimization and case outcomes.

This concept, while important at the first stage of data collection, is covered in Chapter IV.

Timeliness

Even when data meet high standards of quality, they will become less and less useful if not obtained and processed in a timely manner. Like many social phenomena, trafficking in persons and its perpetrators both evolve and adapt quickly. A regular supply of current data is needed to obtain up-to-date insights that will help combat the crime.

Ethical collection

Finally, it is crucial that data are collected and processed ethically, in accordance with data protection principles, and in a manner that upholds the rights of data subjects. As explained earlier, the legal basis for the sourcing of data for evidence purposes has implications for data management throughout the entire data life cycle. Ethical data collection that follows data protection principles is an important part of that.

[20] The actual intervals will vary based on capacity.

It is important to keep in mind that a data point is always more than just a kilobyte or numerical value. Each data point may be a piece of sensitive information offered by an individual willing to put themselves at risk to help combat trafficking in persons. While upholding certain ethical standards is an obvious moral imperative, some ethical principles, such as informed consent and legitimate purpose of use, will also determine the legal basis on which an agency can process and use data (and for what purposes).

Ensuring that data protection and ethics are upheld requires the establishment of, and investment in, proper standard operating procedures for data-producing agencies before any data-collection activities take place. It is these procedures – rather than the actual data – that must be regularly monitored to ensure data protection.

Meeting all of the above properties and their related data requirements requires a substantial, long-term effort from governments in terms of data collection. Many of the more advanced data requirements are rarely met even by those countries most advanced in the process of data collection on trafficking in persons. They are nonetheless all realistic and can be achieved over time with effort and planning, and the aid of the guidelines laid out in this manual.

Using the ICS-TIP to produce/source interoperable data with standardized indicators

Historically, it has been challenging to use administrative data to compare trafficking patterns and forms between countries, because of differences in definitions, procedures, recording and other factors.[21] The importance of standardizing indicator definitions in order to allow for such comparisons cannot be overstated. The advantages are many, but the most important point to stress is that none of the stated objectives of collecting data on trafficking in persons can be met unless this issue is first addressed. Inaccurate data are at best unusable and at worst, lead to faulty assumptions, limp progress and misdirected resources.

What are international classification standards?

International classification standards are essentially instructions for the measurement of standardized statistical concepts in a particular field. They are used to produce national statistics that can be harmonized for international cooperation, comparison and reporting. Manuals for classification standards provide precise definitions of concepts for the purposes of data collection. Typically, they also describe the entity or unit that is being measured. They also often provide explanations of definitions and practical examples for the user. These documents can be useful resources for identifying how data should be harmonized, aggregated and presented.

Classic examples of international classification standards lay out an exhaustive list of every type of unit on a particular subject. The UN DESA International Standard of Industrial Classification of All Economic Activities, for example, categorizes every type of industry in the productive economy. Likewise, UNODC established the International Classification of Crime for Statistical Purposes. These examples are of great interest here because they connect explicitly with the types of data needed to measure indicators on trafficking in persons.

[21] Global Migration Group, *Handbook for Improving the Production and Use of Migration Data for Development*, (Global Knowledge Partnership for Migration and Development (KNOMAD)/World Bank, Washington, D.C., 2017), p. 177.

The major difference, however, is that, in the case of trafficking in persons, multiple types of information are needed in different subject areas for indicators. For example, data are needed not only on the TIP event (a crime), but also on the purpose of trafficking (which may be labour exploitation, in which case the industry would need to be recorded) and whether the individual victimized was trafficked internationally (migration information).

The ICS-TIP provides the standardized measurement classifications for primary indicators, most importantly, the main trafficking in persons event classification, precisely defined based on international legal standards. Following the classification guidance in the manual and adapting national and local data collection protocols to the ICS-TIP will facilitate national reporting by fostering data that are more consistent. The classifications align with the well-established, associated classifications of crime and industry indicators mentioned above,[22] the data for which are likely already collected for international reporting. Using the ICS-TIP to inform locally established data-collection procedures is the first step towards resolving the problem of incompatible data based on differences in definitions of trafficking in persons.

Technical aspects of the ICS-TIP

The ICS-TIP was developed to facilitate the production of national statistics that can be used to improve the national and international evidence base, and in turn to inform policymaking. It provides standardized definitions of core TIP attributes and recommends what data to collect on these attributes based on government capacity to source quality data. It also describes the data format, including the units of analysis and levels of measurement each attribute requires.

Collecting data in the right format and for the right units of analysis

The structure and format of data at the collection stage have real implications for both data harmonization and data storage later in the data life cycle. The central unit of classification used in the ICS-TIP is the TIP event. Similar to the UNODC International Classification of Crime for Statistical Purposes, which relies on the criminal offence as the central unit, trafficking in persons is centred on the event, act or process of trafficking in persons, as defined in the United Nations Trafficking in Persons Protocol.

While the unit of classification (the TIP event) is the central unit, the nature of TIP administrative data means that information should also be collected on the victims, perpetrators and reporting entities involved. Indeed, data on the TIP event itself and the individuals affected may come from different sources. In addition, different types of TIP data are of value to different data consumers for different purposes. Individual data are collected on victims and perpetrators to capture the individual's characteristics, including background and experiences. They can be used by governments, researchers or CSOs to improve prevention and victim protection, and to help law enforcement agencies understand how to target deterrence and other interventions. The process, or event, of trafficking in persons and all of its characteristics are important for reporting crime statistics and identifying patterns, such as the industry of exploitation or what the recruitment process looks like generally. Reporting entity-level data is collected to obtain information on the organization that produces the administrative record and identifies the event as an instance of trafficking, and where possible on the referral of the victim. The three categories of

[22] Specifically, UNODC's International Classification of Crime for Statistical Purposes (ICCS) – Version 1.0 and UN DESA's International Standard of Industrial Classification of All Economic Activities (ISIC) – Rev. 4.

"victim", "perpetrator" and "reporting entity"[23] are therefore units of description in the ICS-TIP, in addition to "event", which is the central unit of classification.

Attributes that further describe the event itself (such as the time and location) and the victim, perpetrator and reporting entity, are also collected. These attributes are treated as disaggregating variables, in line with the UNODC International Classification of Crime for Statistical Purposes.

Figure 6 synthesizes the relationships between the central unit of classification (event), units of description (victim, perpetrator, reporting entity) and disaggregating variables.

Figure 6. The framework of the ICS-TIP

Note: The framework provides the unique identifiers needed to connect the details of a case, but also to break them down for operational purposes. Orange designates the primary unit of classification – the event. Dark blue refers to the units of description. Dotted lines connect the reporting entity to the other three units, indicating that they may disaggregate data by event (central unit of classification) and/or victim and perpetrator (units of description). Disaggregating variables, in yellow, can provide further details and attributes on each of the units.

Part of the added value of using different units and a relational data model like the one described in Figure 6 lies in the way the data can be stored and updated: trafficking in persons can occur over extended periods and involve multiple individuals (victims and perpetrators) and various governmental and non-governmental services (reporting entities). This enables data users to build a more complete picture of the event.

This feature of the ICS-TIP also provides the flexibility needed for it to be used by a wide range of organizations with different processes and purposes, and in different contexts. The specific data model that is most useful and intuitive for a given organization producing administrative data in a national context will depend on that organization's mandate, operational role or focus. For example, in some cases administrative data may only be recorded at the perpetrator or victim unit level and the only available information may be that which further describes the victims or the perpetrators.

Collecting data on the right disaggregating variables

In addition to reporting on standardized attributes of the TIP event, it is important to collect data on several other attributes, or disaggregating variables, in order to enhance

[23] The term "reporting entities" refers to organizations that assist victims, collect data and are otherwise involved in the TIP event.

understanding of the issue and, ultimately, better address it. The data objectives and types of data needed to meet them listed in Table 1 above are a tall order for most, if not all, governments. While many governments are interested in sourcing the best possible data to deal with the problem effectively, the consultation process clearly revealed that some indicators are considered "must have" for administrative purposes, while others are considered "nice to have" if they can be reasonably obtained.

The list of indicators recommended by the ICS-TIP was drawn up on the basis of long- and short-term government needs and the capacity to meet them. They are grouped in steps.

Step 1, the lowest step, is a basic list of disaggregating variables that includes basic information on the primary source of data collection and the core demographic information on perpetrators and victims.[24] Most governments will have to work up to Step 1 data collection in terms of resources and coordination.

Step 2 recommends a series of additional disaggregating variables useful for obtaining more in-depth information on the crime itself and on the individuals involved. The information on the act, means and purpose of the TIP event required in this step is intended to improve government understanding of the way the crime is manifesting nationally. More information is also requested from victims and perpetrators, to gain a better sense of assistance and law enforcement needs.

Step 3 essentially introduces an additional, optional level of granularity in respect of the indicators required in Step 2, with additional information requested on the victim, perpetrator, trafficking event and reporting entity. It is not recommended for most countries, as this level of data collection is likely not feasible. However, it could guide a more in-depth data-collection effort for a subset of cases.

ETHICAL CONSIDERATIONS AND DATA PROTECTION PRINCIPLES

As outlined earlier in the chapter, and beyond their intrinsic value, ethical data collection and compliance with data protection principles play a major part in allowing the use of the data for evidence purposes. It is therefore essential to know the regional, national and/or local legal frameworks and data protection frameworks specific to the particular environment where the data are held. Any data-collection protocols should comply with these frameworks, at every step of the way.

In addition, the principles set out below constitute a baseline for understanding good practices in data-collection ethics and support the development of data-collection protocols that central agencies may work with data-producing agencies to implement.

In many respects, context-specific legal frameworks will determine the rights of data subjects and how their information is protected. Even if legal frameworks in a given context may not fully embed some of these principles, it remains good practice to follow them and develop guidelines based on examples of good practice and legislation in other contexts, such as the European Union GDPR, Australia's Privacy Act and Mexico's Data Protection Law.

[24] Step 1 contains fewer indicators than are indicated in IOM, *Guidelines for the Collection of Data on Trafficking in Human Beings, Including Comparable Indicators* (Vienna, 2009). While the additional indicators are clearly useful for gaining a better understanding of human trafficking, experts considered that the full list, which was designed for data collection in European Union countries, would place too great a strain on governments new to data collection in this field.

Do no harm

At the core of all ethical data collection, regardless of the topic or individual circumstances, is the principle of "do no harm". Indeed, it is so vital to prevent harm to any human subject involved in the data-collection process, at any stage in the data life cycle, that data collection should not proceed, regardless of its merits, if harm cannot be prevented.[25]

While the do-no-harm principle obviously applies to all data subjects in any data-collection or -use scenario, its scrupulous application is particularly crucial when it comes to sensitive data on vulnerable populations, such as TIP data. Furthermore, the principle applies to all stakeholders, not just data subjects, including data collectors, enumerators, partners and researchers.

The potential for harm can crop up in many ways in the data-collection process. To guarantee that data sourced from external providers has been collected ethically, it is essential to ensure that the core elements listed in the rest of this section have been thoroughly covered by first-hand data collectors. This means establishing criteria for risk assessment and laying out the steps for mitigating any risk that may crop up during data collection. The IOM Displacement Tracking Matrix (which focuses on research and data on internally displaced persons) has developed helpful tools for the initial assessment of risk during data collection. It uses the following checklist[26] to determine sources of potential harm.

- Can collecting the data do harm?
- How likely is it that asking this question puts enumerators, key informants, displaced people, the host community or others in (increased) danger/at (higher) risk? What are these risks/dangers?
- How likely is it that asking this question puts the organization and its capacity to carry out its activities in (increased) danger/at (higher) risk? What are these risks/dangers?
- Are there accessible and safe services to support community members if the question triggers the sharing of information on incidents of violence and abuse? Are the enumerators able to refer to these services?

This issue, and the principles set out below, are revisited throughout the guidance manual, as similar ethical standards will need to be set for each stage of the data life cycle.

Specified and legitimate purpose

Another principle at the core of ethical data collection requires that data must be collected (and processed) for a "specified and legitimate purpose". This means that data must serve specific information needs (identified before data collection) and be used for reasons that are deemed necessary and reasonable. As data subjects undergo at least some level of personal risk (even after safeguards have been put in place to mitigate risk), it is essential that the purpose of data collection be proportional.[27]

This legitimate purpose is considered "specified" when it is clearly identified and communicated, especially during the process of obtaining consent (see below). Communicating the specified

[25] See Protection Information Management (PIM), PIM Principles, 5 February 2017; IOM, *Counter-trafficking in Emergencies* (see footnote 8).
[26] Displacement Tracing Matrix, How Can We Do No Harm When Collecting, Storing, Sharing and Analysing Data? (1 January 2020).
[27] IOM, *IOM Data Protection Manual* (Geneva, 2010).

and legitimate purpose of data collected from sensitive or vulnerable populations is critical to the process of informed consent set out below.

Consent

Obtaining the consent of individuals providing sensitive personal information is a foundational ethical data-collection principle.[28] Consent is not, however, necessary for all types of administrative data or for all the purposes for which it can be processed (for example, certain types of data on perpetrators; see also the sections "Ensuring the legal basis for using data downstream is planned for in upstream primary data collection" (above) and "Legal basis for processing, sharing or publishing data" (below)).

The key basic elements of informed and active consent are as follows:[29]

1. All specified and related purposes (i.e. purposes that aim to fulfil the original specified purposes) of data collection are clearly stated;
2. All processes in the data life cycle are disclosed;
3. Access, correction and complaint procedures are explained;
4. All foreseeable disclosures to third parties are disclosed.

Given the specific nature of the vulnerable groups and individuals at risk, or identified as likely victims, of trafficking in persons, more detailed criteria must be designed to meet the ethical standards for this type of data collection.

Of course, consent must be freely given, which means that trafficking victims in particular must be fully aware that they will benefit from services regardless of whether or not they consent to provide data. The information listed above should also be communicated to data subjects in a language they understand, and in a clear, understandable manner.

Furthermore, the notion of "trauma-informed" consent must be at the core of all data-procurement practices.[30] Trauma-informed consent is informed active consent tailored to address the specific needs and account for the particular vulnerabilities experienced by those victimized by trafficking in persons in all of its forms.[31]

Protection of privacy

Protection of privacy is also, of course, central to data ethics, especially in the case of highly vulnerable populations. The guidance on appropriate protection of sensitive data often focuses on the removal of personal information, a process that is described in Chapter V. That being said, it is equally important to establish protocols to protect privacy during the data-collection phase, especially if there is any risk that survivors will be re-victimized on contact with front-line assistance workers.[32] Protecting the identity of survivors is the

---

[28] According to the *IOM Data Protection Manual* (see footnote 26), "Consent must be obtained at the time of collection or as soon as it is reasonably practical thereafter, and the condition and legal capacity of certain vulnerable groups and individuals should always be taken into account. If exceptional circumstances hinder the achievement of consent, the data controller should, at a minimum, ensure that the data subject has sufficient knowledge to understand and appreciate the specified purpose(s) for which personal data are collected and processed."

[29] These elements are derived from the *IOM Data Protection Manual* (see footnote 26).

[30] Brunner, *Getting to Good Human Trafficking Data* (see footnote 6).

[31] Although trauma-informed consent practices are typically established to guide survivor assistance and after-care, trauma-informed consent can (or should) also involve sensitivity to and protection of data collectors and researchers working with data that can cause trauma.

[32] See, for instance, Principle 6 of the *IOM Data Protection Manual* (see footnote 26), on confidentiality: "Confidentiality of personal data must be respected and applied **at all stages of data collection and data processing**, and should be guaranteed in writing" (emphasis added). On trafficking specifically, see International Centre for Migration Policy Development, *Anti-Trafficking Data Collection and Information Management in the European Union – a Handbook* (2007), section 5.2.1.

subject of core guidelines of the United Nations Recommended Principles and Guidelines on Human Rights and Trafficking, which call on States to:

> ensure that trafficked persons are effectively protected from harm, threats or intimidation by traffickers and associated persons. To this end, there should be no public disclosure of the identity of trafficking victims and their privacy should be respected and protected to the extent possible, while taking into account the right of any accused person to a fair trial. Trafficked persons should be given full warning, in advance, of the difficulties inherent in protecting identities and should not be given false or unrealistic expectations regarding the capacities of law enforcement agencies in this regard. (Guideline 6.6)
>
> …
>
> protect, as appropriate, the privacy and identity of child victims and taking measures to avoid the dissemination of information that could lead to their identification. (Guideline 8.9)

ENCOURAGING SHARING OF ADMINISTRATIVE DATA BY FOSTERING AN EQUITABLE, MULTISTAKEHOLDER DATA ECOSYSTEM

There is no one-size-fits-all, standard approach to sourcing and compiling TIP administrative data effectively. The contexts in which data are collected vary widely, with respect to the kinds of administrative data produced, the kinds of agencies producing them, what governments can require from data holders and how willing data-producing agencies will be to share data on vulnerable populations.

When building a data ecosystem to allow TIP administrative data assets to be optimally leveraged in the national fight against trafficking in persons, it is important to first ask the very basic question: what do you need to know? Bear in mind that data environments can vary drastically and that data needs evolve and change just as the TIP situation in a country often does. Different government agencies will have different priorities depending on the scale of the problem, how complex the national situation is and whether it is possible to source data that will help.

Central government agencies will need to define their priorities and determine what data they need to meet them. The section "Gathering the right types of data to achieve government objectives" provides guidance and specific examples in this respect. Next, government agencies have to determine which data to secure and where.

Identifying data-producing agencies, their data assets, and other relevant stakeholders

The aim is to use heterogenous sources of administrative data and build up a data ecosystem that can provide insights into the national picture of trafficking in persons. The main challenges in this respect have been described earlier in the chapter and include the capacity of data-collection agencies and the heterogeneity of data sources.

There are various approaches to building a data ecosystem, although the most effective way to begin is to map the network of potential data-producing agencies and data stakeholders.

Mapping the participants

When mapping the participants in the data ecosystem, a good starting point is to assemble a list of larger, known organizations producing administrative data. Once these primary data-producing agencies have been identified, it may be possible, through a consultative process, to (i) locate others in the system, including less visible data-producing partners and potentially those already sourcing and/or using their data; and (ii) identify and describe the data assets that organizations have. This can be done by surveying partner (or known) CSOs using a snowball sampling method.

Stakeholders other than data producing agencies are also relevant to the data ecosystem. By identifying and mapping key data beneficiaries, intermediaries, potential users, and so on, it is easier to develop a sense of how the data assets will need to flow and where barriers may crop up.

Mapping formal and informal value exchanges

Earlier in the chapter, it was recommended that governments identify and seek data to fulfill specific objectives and needs. When developing a data ecosystem, it is also important to identify the data needs and objectives of others involved in the process and how they may be complimentary, so as to make mutually beneficial arrangements between stakeholders (see Box 2). It is useful to consider the interests of diverse stakeholders with an eye to the formal and informal, or "soft", value exchanges that may exist between them.

In terms of formal value exchange, or the more tangible assets that can flow between partners in the data ecosystem, the most important and sought-after assets are data sets. Other tangible items to bear in mind are relevant forms of data documentation, metadata and any data licences and certificates that may be needed. It is also important to map the monetary cost of obtaining data assets. "Soft" value exchange items that it may be useful to map are partner insights and knowledge, in order to improve service provision or policymaking, and support and advice that can work to improve the data ecosystem as a whole.

Box 2. Data-mapping exercise

This is how to plan the mapping process:

- What are the data assets needed (for what specific purpose)?
- Who are the data stewards?
- What are the existing data-sharing arrangements?
- Where are the opportunities for added value?
- What are the barriers to accessing/sourcing the data?
- How will those barriers be addressed/overcome?

Establishing trust with data-producing agencies

Whether a data-sourcing agency is in the process of establishing new data partnerships or has already been involved in data exchange with data-producing partners, there are clear ways to establish, maintain or improve data-sharing/sourcing relationships. According to experts that have long-term experience facilitating such relationships, the key to successful partnerships is trust.

Establishing trust among stakeholders seems an obvious and straightforward goal, but when it comes to sharing sensitive administrative data, including on vulnerable populations, data-producing organizations can sometimes have conflicting goals and interests. While governments have a range of needs and objectives served by administrative data, for these agencies, data collection is a record-keeping function used to better serve the individuals being assisted or to support judicial proceedings. For example, when recording information on a survivor's personal details and experiences, front-line care providers must ensure that protection of the individual offering the data extends to the data assets as well. Front-line agencies do not want to breach that trust and therefore, understandably, exercise great caution when it comes to sharing data, even with government agencies.

What is needed is a way to reassure data-producing agencies that the people in their care are protected and that their data will never be used for any purpose other than that explicitly stated. This level of trust can be fostered through good data governance. Data governance frameworks, their scope and design are discussed at length in Chapter IV. The focus in this section is on how data governance can support processes that are inclusive and mutually beneficial while fostering trust. In this respect, common practices that can be cited here are formalizing data-sharing agreements or establishing other institutionalized coordination mechanisms (see Box 3). On a more informal level, trust between stakeholders can be built through consistency and transparency, by always adhering to commitments between stakeholders and ensuring that any data sharing, use or publication is reviewed and agreed to by all.

Beyond establishing trust, partnerships can also be strengthened through equitable data-sharing arrangements. Creating mutually beneficial relationships can ensure that the formal and informal value exchanges serve the interests of data-producing agencies as well as those of the government agencies sourcing the data. It is also important to ensure that data-producing agencies are protected, that decision-making on data matters is inclusive, and that data ownership is well-defined and respected.

In Canada, the Statistics Act[33] provides the national statistical office, Statistics Canada, with a legal mandate and establishes the legal basis for obtaining administrative data from governmental and non-governmental agencies. When it comes to TIP administrative data, Statistics Canada has also spent decades building relationships with primary data-producing CSOs and other data-holding agencies, cultivating a data-sharing culture that eases the burden and tension that can arise when protection organizations are requested to provide sensitive data. Beyond general data governance processes (see Chapter IV), this has included the establishment of the Ethics Committee, which operates under a necessity and proportionality framework, and seeking the agreement of data providers for each purpose for which the data are used.

Another novel approach to fostering trust is to establish a quasi-autonomous, independent organization to act as a clearing house, or data trust, that has full stewardship of the data, including personal and sensitive information. This organization may be linked to, or part of, the national rapporteur's office or other similar entity. An example of such an "independent clearing house" is the NGO CoMensha in the Kingdom of the Netherlands (see Box 4). CoMensha is an independent organization mandated to source and manage data separately from the government. It then reports the data in aggregate form to the national rapporteur. Generally, by sourcing and processing sensitive data from CSOs before passing them on

[33] Available at https://laws-lois.justice.gc.ca/eng/acts/s-19/fulltext.html.

to the government, the independent data trust serves as an added layer of protection and safeguards against the use of data assets for any purpose beyond those agreed with the data subjects and primary data-producing agencies.

Box 3. Legislating for reporting: some examples

One possible route to obtaining valuable information on trafficking in persons is to make it a legal requirement for front-line and other data-producing agencies to provide data to government data-sourcing agencies. There are some notable examples of this process working to produce a more robust data ecosystem that benefits the development of data-driven policies and programming.

In Brazil, for instance, the Federal Public Labour Prosecutor's Office can now source administrative data from several government agencies thanks to a legislative requirement intended to break down the barriers between agencies and enhance data-sharing. Data sets are seamlessly linked through the use of encrypted identities. Technically, the data are held in a "data lake" and can be analysed in raw form without extracting them or exposing sensitive information. Previously the data assets held by various agencies existed in silos and were less commonly shared.

In 2015, the United Kingdom introduced the "duty to notify". Designated first responders (local authorities, police forces, CSOs and government agencies) usually refer potential victims to the Home Office Single Competent Authority as part of the national referral mechanism. However, potential adult victims must consent to be referred to the mechanism. If they do not consent (because they wish to remain anonymous, for instance), there is still a duty to notify the Home Office.[a] In this process, "first responders do not collect the potential victim's personal details" but may collect some information on the crime, including, for instance, nationality or types of exploitation.[b]

The experts consulted nevertheless caution against the introduction of a blanket approach mandating CSOs, for example, to provide data, especially without first considering the context. Such a mandate might in some cases do more harm than good if the CSOs are unable to meet the requirement because of weak administrative capacity or unstable funding or if they have limited trust that the data assets will be used only for purposes that serve the data subjects' interests.

[a] More information on the duty to notify and the referral mechanism is available at the United Kingdom Government website.

[b] See United Kingdom Home Office, Official Statistics – Modern Slavery: National Referral Mechanism and Duty to Notify statistics UK, Quarter 1 2021 – January to March, second edition (updated 27 January 2022).

Fostering coordination and collaboration between organizations on the same administrative records

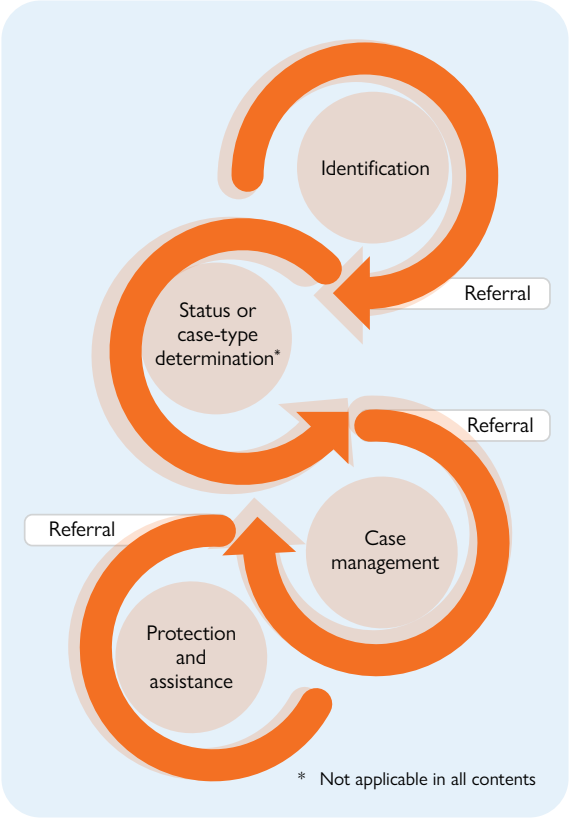
Coordination between counter-trafficking entities during actual operations to address a specific case of trafficking can pose many challenges. Some relate to the sharing of data and information. In turn, poor coordination among data-producing agencies makes it difficult to meet data-collection and -sourcing priorities and is a regular barrier to the collection of high-quality administrative data.

To help tackle coordination issues, it is key to develop and implement a comprehensive data governance framework involving the organizations responsible for collecting data and the institutions that house and analyse them. Chapter IV is devoted to general issues of data governance. By contrast, this section looks at some specific examples for coordinating the input of multiple agencies on the same administrative record, leveraging existing referral and protection mechanisms.

The main structure for achieving coordination of protection services is through a national referral mechanism, which is a process of cooperation between multiple stakeholders to provide protection and assistance services to victims of trafficking. The process includes the various components or steps for providing protection and assistance. These steps may vary in each country, but they generally include identification, status or case-type determination, case management and the provision of protection and assistance services.[34] National referral mechanisms are an invaluable tool in the fight to prevent trafficking in persons and assist the victims. Their development and implementation can greatly improve the provision of protection and assistance for those vulnerable to exploitation or identified as being in situations of exploitation.

Such coordination systems are necessary because victims of trafficking have a wide array of needs that cut across sectors and providers, and it is unlikely that any one government entity or organization can meet them all. Multiple and overlapping protection systems might exist in a specific context, with multiple organizations, each with a different mandate, providing different services.[35] Figure 7 illustrates the general process for setting up a system of contact points to facilitate the care process.

Figure 7. Process for setting up a system of contact points[36]



The sections below describe two institutional ways of improving coordination and ensuring data can be linked across organizations within the national referral mechanism.

[34] *IOM Guidance on Referral Mechanisms for the Protection and Assistance of Migrants Vulnerable to Violence, Exploitation and Abuse and Victims of Trafficking* (IOM, Geneva, 2019), p. 7.
[35] Ibid.
[36] Ibid.

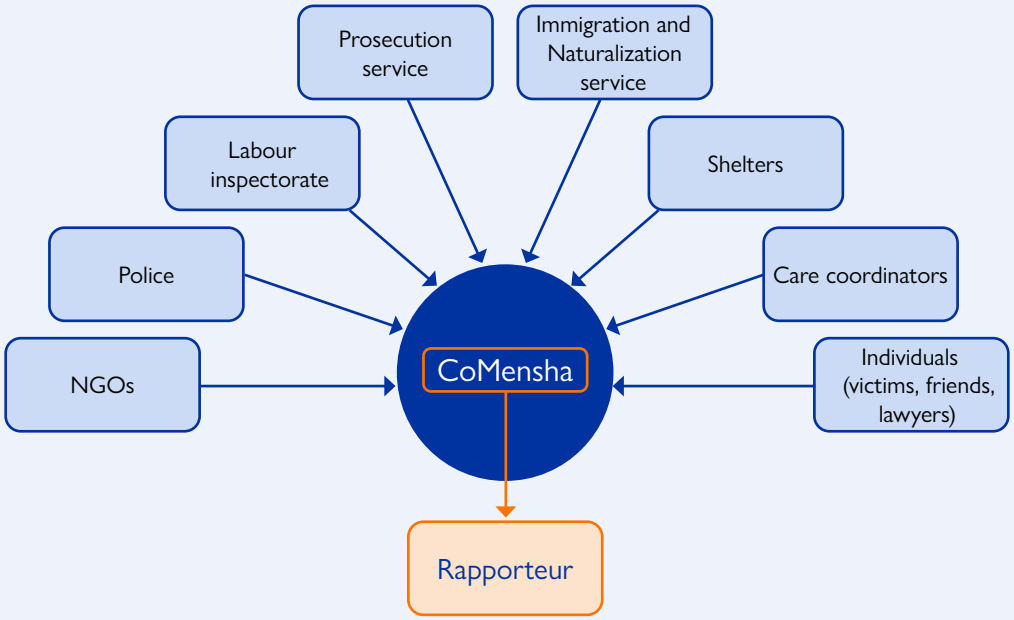
A central agency acting as a data trust and independent clearing house

One novel approach to harmonizing data sets by linking personal information is to establish a quasi-autonomous, independent organization to act as a clearing house, or data trust, that has full stewardship of data, including personal and sensitive information. In particular, this type of clearing house can be tasked with managing referrals, coordinating the input of data and linking data on the same administrative record through a system of unique identifiers. An example of this model is CoMensha, in the Kingdom of the Netherlands (see Box 4).

Box 4. CoMensha

CoMensha is an independent clearing house or data trust, a "legal structure that provides independent stewardship of data"[a] (see Chapter IV for a discussion of data stewardship). This is a system whereby independent organizational trustees source and manage data separately from a government agency. CoMensha manages the data in the interests of the trust's beneficiaries, ensuring that data assets are only used for the purposes specified by the data suppliers. It then reports the data in aggregate form to the national rapporteur.

Figure 8. How CoMensha works[b]



CoMensha has assembled a rich inventory of data assets from a broad range of front-line agencies producing administrative data (see Figure 8. How CoMensha works) on TIP cases, including highly sensitive data from victims. While this is also true of many governments, the independent data trust model allows CoMensha to overcome some of the challenges faced by governments, sourcing data so that all the records collected by the clearinghouse can be linked with personal identifiers across the diverse data sources.

[a] OECD, *The Path to Becoming a Data-Driven Public Sector*, OECD Digital Government Studies (OECD Publishing, Paris, 2019).

[b] The figure is taken from the workshop presentation of the Dutch National Rapporteur on Trafficking in Human Beings and Sexual Violence against Children.

A shared information management system hosted by the central agency

Importantly for the purpose of this guidance, it is possible to establish an information management system to support the activities of the national referral mechanism, hosted and maintained by the lead agency (or one of the agencies that is part of it). This is the case, for example, in the United Kingdom, with the Home Office being the agency hosting the information management system. This can considerably facilitate data sourcing, particularly with respect to the issue of overlapping coverage. Indeed, organizations that are part of a national referral mechanism collaborate on the same administrative records, with designated authorities initiating the referral process and a defined sequence of authorities updating or amending the records as cases progress. In the absence of a common, inter-agency information management system to support the national referral mechanism, all the organizations involved would each work on their own record of the same case, creating multiple data sets with overlapping coverage. However, where multiple agencies input data to a common information management system hosted by one lead agency acting as custodian, inter-agency data governance arrangements are all the more important to ensure that contributing agencies are clear on what purposes the data they contribute to the system can be used for (see Chapter IV for more details).

It should be clear from the discussion of data-sourcing strategies that the key to developing effective systems lies in building trust and supportive relationships with organizations that serve victims and collect data first-hand, and with other stakeholders at the national level, such as law enforcement/criminal justice agencies. The next section will cover practical ways to bolster capacity for producing high-quality data.

IMPROVING CAPACITY TO GENERATE HIGH-QUALITY ADMINISTRATIVE DATA

Front-line and other agencies producing administrative data are often under-resourced even in the best of cases. Under-resourcing can lead to a host of problems, some of which are more relevant to data-collection capacity: high staff turnover; staff and management lacking technological experience or training; limited technical infrastructure; and lack of specialized staff to support and maintain IT systems. In general, the administrative capacity needed to collect high-quality administrative data is often seriously lacking.

Investing in information management tools

Investments in core capacities are required for data-producing agencies to set up data-collection activities. Such agencies will likely need training, IT equipment, software and information management systems requiring development and/or configuration. These expenses are often not a one-time investment – systems need to be maintained and updated, new staff must be trained as systems evolve, licence fees must be paid, and so on. Whatever is put in place should be sustainable.

To help ensure that this is the case, ahead of these concrete investments, a needs assessment should be conducted to understand the data needs and the capacities of the data-producing agencies. A data management plan must be set up, including definitions of data management roles and protocols. Much of the guidance in Chapter IV will be useful in this respect.

In addition, publicly available material such as the Human Trafficking Case Data Standards, Toolkit and Guidance[37] can also help front-line agencies (and potentially other data-producing agencies) build information management systems more easily. The toolkit, which encompasses the ICS-TIP and standardized fields related to case management, provides tools and guidance for front-line counter-trafficking agencies on the standardized collection, management and potential sharing of information related to TIP cases.

There are also promising examples of large international CSOs, like Liberty Shared, providing smaller, local CSOs in Asia and elsewhere with technological assistance in the form of content management systems and training.[38]

Investing in human resources and capacity

It follows that using new technologies requires time and, importantly, capacity. This could mean investing in training personnel at entry points and/or even new hires.[39] It is likely that such investments will not be one-offs – as technologies evolve and staff turns over, additional investments will be needed.

Training for entry-point staff will likely need to cover the data management plan put in place in the data-producing agency: who has what role and who has access to what data, the data protection measures put into place, how to use the IT equipment and infrastructure, and how the data are to be safely used and shared. The technical transfer of data to the national repository could also be covered.[40]

37 Available from https://github.com/UNMigration/HTCDS.

38 According to the experts consulted, training and technological capacity-building cannot fully resolve issues of under-resourcing unless governments are willing to step in with more resources.

39 IOM, *ASEAN and Trafficking in Persons: Using Data as a Tool to Combat Trafficking in Persons* (Geneva).

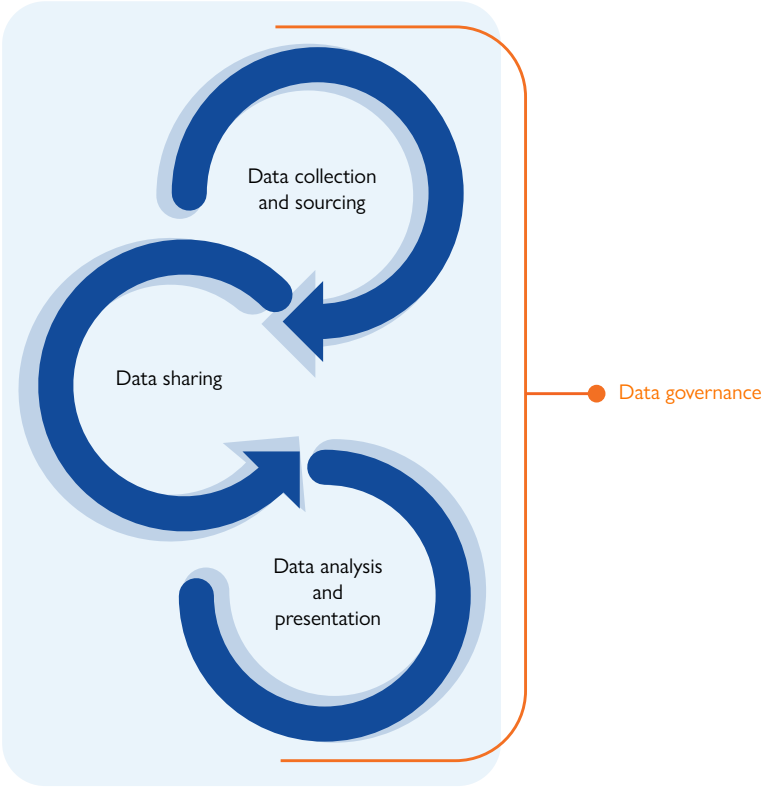
40 International Centre for Migration Policy Development, *Anti-Trafficking Data Collection* (see footnote 31).

CHAPTER 4:
DATA GOVERNANCE

Chapter III focused on the common issues facing data-sourcing agencies and how to facilitate the sourcing of TIP administrative data, from general principles to concrete steps, including recommendations to align with and adopt the ICS-TIP where feasible. It also emphasized the importance of building effective collaboration with organizations that produce administrative data, in order to protect data subjects willing to volunteer sensitive information to enhance understanding of trafficking in persons and help governments better address this crime.

In this chapter, the focus is on the development of data governance frameworks (see Figure 9) for the secure and efficient management of multiple administrative sources of data through a process for decision-making and accountability that is inclusive of all stakeholders. The chapter starts by outlining some of the general objectives that a data governance framework should fulfil, from the process of decision-making and establishing accountability to fostering inclusivity and effective coordination among all stakeholders. It then goes through the roles to be created and their respective responsibility in maintaining and enforcing the framework, followed by the rules to be put in place. The final section looks at roles and rules in the specific inter-agency context, which is fairly common in the realm of administrative data on trafficking in persons.

Figure 9. The data life cycle: data management



The chapter is designed to be useful for officials at different stages of the development of data governance frameworks. For officials setting up an entirely new system or repository outside a particular government agency, it provides instructions for getting started. For the many government agencies already applying data governance standards within broader legal data protection frameworks, it helps enhance understanding of how to apply the rules for the management of data on trafficking in persons specifically, which may present unique challenges in terms of security and organization. Some of the actions suggested may help improve existing governance frameworks and data management strategies.

GENERAL OBJECTIVES OF DATA GOVERNANCE FRAMEWORKS AND PROCEDURES

Data governance is essentially like governance in any other domain: it entails establishing policies and ensuring their consistent and effective implementation. To support this process, a data governance framework is needed to define clear procedures, rules, roles, and responsibilities to ensure consistency and accountability. Importantly, this includes a process for decision-making that is inclusive of all relevant stakeholders.

Data governance, in the context of this manual, applies to the process of codifying the rules and procedures required for appropriate and robust management of TIP administrative data assets at the national level. It includes a framework of decision rights and processes of accountability assigned to the officials managing the data at every step of the process. Any data governance framework should address the following considerations:

- Defining data management roles and the scope of data management rules;
- Determining who has decision-making authority and responsibility for different aspects of data management, including access, sharing and use;
- Determining who has rights over the data and derivatives;
- Determining how data assets should be used;
- Determining who must be consulted and who must be informed regarding decisions over data management and use, and where prior approval and agreement must be sought;
- Providing for dispute resolution mechanisms, including procedures to manage disputes that cannot be resolved and result in an agency exiting from the agreement;
- Establishing a framework for risk management and mitigation;
- Establishing lines of accountability.

It is commonplace to think of data security as something very technical involving high-tech cloud storage, cutting-edge advances like blockchain or sophisticated types of encryption software. Data security tends to be thought of only when one of these techniques or technologies fails or falls victim to a cyberattack, leading to the high-profile exposure of private information. What is not as commonly understood or discussed is how data governance frameworks can help to establish the straightforward protocols needed to secure private information. The roles and rules established by a robust data governance framework create the systems and accountabilities needed to protect data assets even in the face of technical failures.

Accountability

A primary aim of data governance is to create the kind of accountability that can establish public trust in the protection of administrative data assets at the national level. According to the United Nations Secretary-General, data protection and data privacy should be thought of as digital human rights:

> Effective personal data protection and the protection of the right to privacy in line with internationally agreed standards are imperative. Human rights-based domestic laws and practices for the protection of data privacy, including enforcement mechanisms such as access to judicial review, or fully independent and well-resourced data protection authorities, are needed to address the use of data by private companies or Governments.[41]

[41] United Nations, Road map for digital cooperation: implementation of the recommendations of the High-level Panel on Digital Cooperation, report of the Secretary-General (A/74/821 of 29 May 2020), para. 44.

While this statement reflects growing concern at the complete saturation of our daily lives by the digital world, it is certainly not the first time that data privacy has been recognized as essential to fundamental rights and freedoms[42] and likely not the last.

Accountability to data subjects, data-producing agencies and end users of government data (whether internal or external) through the protection of data assets is established through clearly delineated responsibilities to uphold the rules of the data-sourcing, management and -sharing processes.

Delimited purpose

Clearly defining and limiting the purposes for which data can be used is particularly important in the case of TIP administrative data, particularly in an inter-agency environment. What if, for example, the data collected to assist and protect victims of trafficking are used to inform and target immigration and border control measures? Would this mean that some of the victims coming forward would risk deportation? This would clearly discourage them from doing so and could give traffickers more leverage – hence the need to "firewall" TIP data from other policy areas.

In another example, what would happen if victim data that were to be used to investigate perpetrators (a purpose for which the police force collecting the data may have obtained the requisite consent and permissions) were also used for prosecution without the victims' explicit consent? This would be a serious violation of the data subjects' trust and a threat to their safety.

Dealing with the heterogeneity of data assets

Different rules will be needed to facilitate the management, storage and eventual usage of diverse data types. Different data types can and will be used for different purposes. Data governance frameworks must therefore allow for and facilitate this diversity.

It is unlikely that one set of criteria for data management, storage and security will apply in the same way to all types of data in any administrative data governance plan, especially when it comes to the indicators suggested by the ICS-TIP, which call for data to be collected on different types of individuals, industries, criminal acts, organizations, relationships, and so on. It is important to think about the varying uses and protection requirements of different types of data and to remember that both the usefulness and sensitivity of one type of data can change when linked with other types of data.

For example, in terms of data protection, national laws tend to require less protection for personal information on perpetrators, at least after prosecution, than for sensitive data on victims.[43] In fact, in some cases, the law places perpetrators' identifying information in

[42] See notably the right to respect for the private and family life of individuals set out in the European Convention on Human Rights and the right to protection of personal data set out in Article 8 of the Charter of Fundamental Rights of the European Union. See also the general principles of Community law, cited in International Centre for Migration Policy Development, *Anti-Trafficking Data Collection* (see footnote 31).

[43] For example, the European Union GDPR "does not apply to data processed by public authorities in the course of the prevention, detection, investigation, and prosecution of crime, which is governed by Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 'on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data'".

the public domain with little, if any, protection, as evidenced by the appearance of national offender registries (especially on convicted sex offenders) in many countries.[44] However, consider the implications of connecting perpetrator data with that of the victims involved in the same trafficking case. Linking such data can be extremely useful for prosecution, compensation and assistance but extremely dangerous in terms of the potential to violate the victims' right to privacy.

Finally, the rules set must be able to consider new types of data not yet sourced. As new technologies are developed to collect data (satellite imagery or mobile phone technologies), more data will become available to use alongside the core ICS-TIP indicators. Governance frameworks should, at a minimum, be developed with an eye to novel developments in data sourcing.[45]

Dealing with risk and data classification

Managing diverse data sources, particularly those with sensitive information on highly vulnerable populations, requires sharp attention to potential sources of risk. This is where the concept of risk classification and user-based authorization becomes important. Privacy rules must be set according to varying levels of risk to data subjects and data assets. A classification system must be developed that establishes risk levels for each data type and the relevant access controls (see Table 2 below). Highly sensitive information on victims of trafficking in persons, especially the kind that could lead to individuals being identified and/or retraumatized, must always be protected at the highest level of security.

At the same time, ethical, safety and legal considerations must be balanced against the need for data and evidence. Data can be managed as an asset for the benefit of all counter-trafficking stakeholders. Those stakeholders can also be involved in governance of data assets to help ensure that this is the case. Fortunately, there are many ways to handle data that protect data subjects and offer the information needed for policy development and research. Chapter V describes some of these solutions in detail.

Clear and comprehensive organization of data assets

One fairly obvious principle of data governance and management is that data must be properly organized. Clear protocols on where to store data safely, how they will be formatted and accessed, how long they are relevant (or may be held legally) and what to do with them when they are no longer relevant will go a long way towards thwarting risks to security and quality.

Many of the rules in the data governance framework will deal with these practical issues of data organization. The section on administrative metadata below covers common organizational issues, but there are several general ideas to keep in mind when deciding on storage/maintenance rules, to ensure that the data are FAIR. The FAIR data principles offer the guidance needed to establish data governance policy for the most efficient data

[44] See also, for example, the detailed information on perpetrators figuring in the case history narratives featured in the UNODC SHERLOC database.

[45] One notable source of guidance is a November 2017 United Nations Development Group document, entitled Data privacy, ethics and protection: Guidance note on big data for achievement of the 2030 Agenda. Another manual published in 2020 by the World Bank, *Data Collection in Fragile States. Innovations from Africa and Beyond*, also discusses many innovative approaches to working with big data.

structure and storage by any organization, agency or department.[46] They are intended to ensure that data are:

Findable:	Data must be uniquely and persistently identifiable (that is, systematically organized in a manner that allows specific data to be easily located or recalled).
Accessible:	Data can always be obtained with appropriate authorization.
Interoperable:	Data must work well with, or link to, other relevant systems.
Reusable:	Data should be well described using (descriptive and structural) metadata (detailed below).

Effective coordination

Possibly the most underemphasized aspect of data governance, one that is conspicuously absent from many materials and online resources on the subject, is the role of effective coordination between relevant stakeholders. Given the range of data sources and stakeholders with vested interests in a central repository, the development of a data governance framework for data on trafficking in persons requires real attention to coordination issues. Anticipating coordination challenges before they arise, and establishing the mechanisms to deal with them, will be of immense benefit.

As the previous chapter made clear, effective coordination between data-producing agencies and government agencies sourcing data from them is key. Once the data have been obtained by one or several agencies, streamlining coordination remains a high priority, as it is central to the implementation of data governance rules, especially concerning data ownership and use. Without clear rules on how various parties will work together and who will be accountable for the protection of data at every stage of the process, it will be difficult to establish the trust needed to build buy-in.

Lack of coordination can quickly lead to problems between departments that may have overlapping mandates. Since trafficking in persons crosses over many domains (e.g. women and children's rights, criminal justice, migration, labour), some agencies may be reluctant to cooperate, as there will likely be divergent interests between different stakeholders on the kinds of policies the data can be used to support, for example. It follows that data governance rules must specify not only who is accountable for the data but also who can use the data and how.

ROLES

Establishing public trust in the data management process and technical veracity of the data requires instituting and maintaining a set of data governance rules. Setting and implementing standards for effective oversight of data collection and the proper handling and eventual use of data is the work of a government official or team that has intimate knowledge of the system and processes of the government administrative environment and of the national landscape of trafficking in persons (and related) data. In other words, a key aspect of the rules is to institutionalize the necessary roles.[47]

[46] The FAIR principles are applicable to TIP data when the overarching ethical principle of doing no harm is fulfilled. The related topic of safe data sharing and de-identification is more extensively dealt with in Chapter V.

[47] Kristen Wende, A Model for Data Governance – Organising Accountabilities for Data Quality Management, in *ACIS 2007 Proceedings*, 80 (2007).

Data stewardship

Data stewards bear primary responsibility for the development and implementation of rules on data management (operating within and upholding the data governance framework). They are the official or entity using the ICS-TIP and this manual to establish the rules for sourcing and maintaining data on trafficking in persons. They should also ensure that the framework for data protection standards is built on local and national legislation on the privacy rights of data subjects. If national data protection laws are not up to date or do not provide robust protection to data subjects, the framework can be grounded in well-developed policies from other regional[48] or national[49] legislation or international organizations.[50] Strict guidelines for data protection should be less of a challenge for data-producing agencies such as the police or other government agencies, but may limit the data that can be accepted from CSOs / shelters/partners that have difficulties with compliance.

The role of data steward may appear, but is not always, highly technical. In essence, most officials in charge of protecting information already occupy a position of data stewardship. Moreover, if they can draw on the technical expertise of a team and other experts in a government department (such as a national statistical office), data stewards may not need an extensive technical background to develop effective policies and protocols. They can establish an effective data governance protocol using the guidance offered in this manual.

Data stewards are responsible for developing the data governance rules, detailed later in the chapter, and overseeing their monitoring and implementation. Such rules relate to data access (who can access what data and how), data storage (deciding between cloud and server storage, but also on archiving and disposal), data security protocols and data documentation.[51]

Data stewards may be asked to handle complaints for breach of privacy, data access requests and requests from data subjects for access to or removal of data. They maintain records pertaining to consent, disclosure of data to third parties in any form (see Chapter V), and data retention and disposal.

The data steward's role extends beyond drawing up the management strategy and main data protocols. No matter what type of technical environment houses the data, an individual or team will ultimately have to be accountable for the data assets. Having a dedicated role ensures that the standards set will be upheld and public trust in the process maintained.

However, when it comes to the more technical duties of building and maintaining databases, the agency or organization managing the data will likely need to rely on a data custodian.

---

48 The European Union GDPR is one example.

49 The Canadian Statistics Act, which is described later in this chapter, is a good example. Other potential sources are the governments that have ratified the 1981 Council of Europe Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (the list of ratifications is available at www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=108).

50 Guidelines and principles have been produced by intergovernmental organizations and by the public and private sectors. Examples are the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data and the UNODC Privacy and data protection module on data protection legislation.

51 Specific tasks may include the following: drawing up an overall data management strategy; assigning risk categories to data and establishing protocols for the protection and sharing of each type of data; defining user roles and access levels; setting the rules for data transfers; designating how/where data are stored; maintaining standards for encryption and de-identification; determining data relevance and when to archive/destroy data; and establishing standards for metadata.

Data custodian

The role of data custodians is to oversee the technical aspects of transferring, storing and maintaining data. Other key functions often include setting up and/or maintaining the database infrastructure. Data custodians ensure that data sets meet the required standards of quality and are clean, accurate and timely, and query data for transfer, sharing or publishing at the request of data managers/stewards. Overall, the position is often categorized as an ICT role.

Additional roles

Additional roles may include those of data trustee and data manager(s). These titles and their responsibilities vary between organizations. Data managers tend to be responsible, not so much for data management decision-making as for policy implementation with a view to ensuring the day-to-day functioning of proper data management. The role of data trustee may only be necessary if centralized coordination is required at the top, which could very well be the case when multiple organizations or agencies are collaborating to maintain data collectively. The trustee ensures that the interests of all data stakeholders are met and, when relevant, that multiple data stewards agree on data governance policy.

RULES

Formalizing the rules for establishing the optimal data governance framework via documentation of protocols on access, storage, security and the roles/responsibilities presented above requires first understanding the good governance principles that the agency will uphold. While the specifics may vary based on ownership of the data and legal frameworks, among others, the overall strategy will need to address several basic considerations. These are outlined under "General objectives of data governance frameworks and procedures" above.

Some of the core questions to bear in mind when drafting a data governance framework are discussed in this section, which ends with guidance on how to formally document these procedures in official metadata documentation.

In a complex inter-agency environment, diverse stakeholders may have diverging interests on different aspects of the stewardship of data assets, particularly those they have produced or contributed to. This includes how data are sourced, maintained, shared and used, by whom, and for what purposes. All of these issues will need to be addressed when developing the data governance framework in a process that involves all relevant stakeholders from start to finish, to ensure the level of transparency and trust required for successful collaboration.

Who can access (which) data, how and for what purpose?

Deciding which data to share and how gives rise to many challenges and questions:

- When can the data be considered anonymous (or de-identified to the extent that a data subject cannot be re-identified)?
- Who has (what kind of) rights over the data?
- What can the data be used for? What can they not be used for?
- Who can access the data and in what form?
- How does data subject consent apply to policies of data use and sharing (and in what forms)?

While legal and ethical standards provide guidance on some of these questions, data-sharing protocols vary depending on the content and format of the data, and the associated privacy risks. Taking all of these elements into account will inform decisions about data protection standards and shape the protocols needed to govern the data-sharing and -publishing process for the government agency or data-holding organization.

There are two main ways to go about allowing data access to users outside the primary data-holding agency: to share or to publish. Each has its own implications for privacy. For the purposes of this chapter, they are defined as follows:

> **Share:** securely transfer (requires identifying who can use what data in which form)
> **Publish:** make publicly available (requires de-identifying data for public use)

The manner in which third parties receive data (or the decision to deny them access to data) is determined by many factors, the most important of which is the extent to which sensitive, personal data have been de-identified. De-identification is successful when the data are free of information that is considered "personal" and the individual data subject can no longer be identified by this information. Chapter V provides more details on de-identification and examples of de-identification techniques.

Concretely, the rules need to assign levels of security/risk to each data type, specify who has clearance to access the data and assign roles for implementing procedures. In general, data access should be provided on a need-to-know basis – people have access strictly to what they need to perform their duties or to achieve the purpose for which the data were shared. Risk/benefits assessments can be conducted to facilitate the decision-making process.

Table 2 provides an example of the different classes of data that can be created and the associated risk levels. The most sensitive data are individual data containing directly identifying information, such as names and addresses. Central agencies may never hold this type of data, as such identifiers will likely be removed before the data are sourced. In any case, this is the kind of data that will have to be subject to special security rules, including encryption (discussed below) and limited access. It is important to note that even when directly identifying information is removed, individuals in the data set may still be identifiable (through prior knowledge of victims, unusual victim characteristics, and so on). Other types of data that are highly anonymized, using various methods such as aggregation, may be less sensitive and require a lower level of protection because they pose a minimal risk. More information on classifying risk, de-identification and assigning protection levels is provided in Chapters V and VI.

Table 2. Data classification and risk levels*

Classification level	Description/examples	Risk level	Access controls
Secret	Highly sensitive data containing personal information to be accessed only by a subset of internal users with clearance Example: victims' names, addresses/ locations, telephone numbers, passport numbers	High	Data viewing and modification restricted to users on a strictly need-to-know basis, as determined by data steward

Classification level	Description/examples	Risk level	Access controls
Confidential (internal)	Sensitive data that must remain internal Example: specific location of trafficking routes that may disclose the identity of the victims or sites of assistance to victims	High to moderate	Data viewing and modification restricted to authorized users as needed for specific roles, determined by data steward
Restricted (external)	Data intended for trusted third-party sharing only Example: data without direct identifying information shared with trusted research partners. As little information as possible should be shared to satisfy the research purpose, and even then it may still be possible to identify the victim through prior knowledge or by cross-referencing another data source (see Chapter V).	Moderate	No restriction for trusted third party, under a data-sharing agreement specifying intended use and data-sharing modalities
Public	Data intended for completely unrestricted use Example: safely aggregated data, synthetic data	Very low risk	No restriction

* For more information on the kinds of data that can be shared publicly versus kept private or securely shared with third parties, see Chapter V.

Where to store (active) data

Data are typically stored and managed on either cloud services or on-premises systems (personal computers or local servers). These options have different advantages and disadvantages.

In recent years, the technology sector has invested heavily in cloud software services over on-premises options. As a result, the former often offer more options and functionality than the latter. Many cloud providers have invested heavily in security, including audit, authentication, physical security and operating procedures. Cloud software services are also maintained, updated and tested by the provider, whereas on-premises options are managed and secured by the organization using the system.

Cloud services still use physical data centres, which are owned by the company providing the cloud services – a "third party". An adequate legal/regulatory/security framework needs to be in place to understand how data are protected. Transparency is another requirement, so that it is understood how the company (or another entity) will use any data stored for its own purposes.[52] This becomes more complicated if data centres are in a different jurisdiction from the government responsible for the data, because the country hosting the data centre could requisition the data, possibly without the other government even knowing.

[52] There are a variety of ways cloud companies may use data for their own purposes – some may be acceptable, others not. For example, it might be concerning if a company uses the data to profile authors, target advertising, or simply to sell to third parties. Nevertheless, there are more legitimate uses – for example, it is quite common for companies to track usage data on platforms, which can help guide them on how to improve features and the user interface. These might be anonymized audit data or metadata, and do not include sensitive data.

Some cloud providers offer advanced encryption key options that can mitigate these issues in certain circumstances. Typically a cloud software service will manage production (or live) data services, separate from archive and disaster-recovery services.

Regardless of the option chosen, ensuring a secure network connection (whether local or to the cloud) is paramount. If a database or information management system is to be used by multiple organizations, then these organizations will need to be provided with a level of access to the network that is chosen.

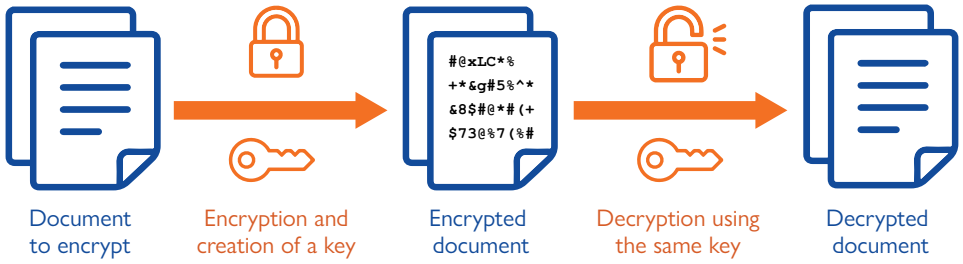
Finally, if any information is to be stored locally on a user's computer at any point, there must be appropriate protections on that device (it should usually not be a personal device) and users should be trained on appropriate data management polices and procedures.

How to keep data secure

In addition to protecting data by storing it in secure locations, other simple steps can be taken to ensure added privacy protection; they have become the standard for international data protection. Advice is widely available on the Internet, from setting up strong passwords to logging off terminals after use.[53] Encryption is another such step, and this is what this subsection will focus on.

Encryption is the process of converting text into incomprehensible code and using a key to protect the original format of the text, as shown in Figure 10.[54] How encryption works is described in Annex 4.

Figure 10. How encryption works



The jargon used to describe the encryption process makes it sound more intimidating than it is. Encryption is simply the process of taking plain text (or numerical indicators) and converting it into a series of unreadable, seemingly random characters. Those entrusted with data protection use a specially assigned key to decrypt or transfer data back into readable text when needed.

Software and encryption issues aside, it is essential to note that cyberbreaches and attacks are most often caused by human error and not by failures of secure encryption standards. In other words, hackers use various techniques to manipulate data protectors into giving out secure information or enabling access. To protect data from human error, data protection policies must include information and training for those with access to confidential data and encryption keys.

---

53 For instance, see the guidance provided by the United Kingdom Information Commissioner's Office (https://ico.org.uk/for-organisations/sme-web-hub/whats-new/blogs/11-practical-ways-to-keep-your-it-systems-safe-and-secure/) or National Cyber Security Centre (www.ncsc.gov.uk/section/advice-guidance/all-topics and www.ncsc.gov.uk/cyberaware/home).

54 IOM, *IOM Data Protection Manual* (see footnote 26), p. 75.

How long to store data: archiving and disposal

Data have a shelf life beyond which their usefulness can diminish. There are a few reasons for this. Older data may not have been collected using the same sampling method or empirical rigour as more current data, making them no longer comparable to more current incoming data; they may have been collected for indicators identified using different or less standardized definitions, also compromising their comparability with more recent information; and they may date back long enough to no longer be useful to describe current phenomena in general.

The determination whether data have exceeded their shelf life depends largely on the relevance of the data to fulfil a specific purpose. That purpose can vary between stakeholders. The original agency or organization collecting the data may have had a specific purpose related to local reporting laws, monitoring and evaluation efforts, and so on, giving the data a certain, sometimes shorter, shelf life. The same data may have value to other agencies, governments or researchers that lasts far beyond this initial cut-off.

Box 5 contains a checklist that can be used to determine when data need to be separated from the main holdings and either preserved for the historical record (i.e. archived) or safely destroyed.

Box 5. Establishing data relevance

- Have inaccuracies affected the quality of data?
- Have any updates and significant changes rendered the original record of data unnecessary?
- To what extent is the original record still capable of adding value to the objectives of the agency/department, and is it worth continued storage?
- Have the data subject's circumstances changed, and do these new factors render the original record obsolete and irrelevant?[a]
- Can "active data" be separated from "inactive data," and has sufficient time elapsed to render the "inactive data" irrelevant?
- Can the irrelevant and unnecessary personal data be used for statistical or research purposes that are compatible with the specified purpose for which they were collected?

Source: IOM, *IOM Data Protection Manual* (Geneva, 2010), p. 75.

[a] This item in particular will likely be very difficult to assess, especially in respect of victim-centred data coming from NGOs. It may be more helpful for data on traffickers and trafficking cases, where additional data are accessible through criminal justice records.

Beyond issues of relevance, some data come with legal requirements that they be either archived or destroyed after a certain time. In cases where data subjects have consented to the use of the data for a specific period, the data must be disposed of once that period has ended if they have any remaining direct identifiers. The terms should be specified in agreements with data subjects and data-producing agencies before the data are sourced.

There are two primary types of inactive data record: permanent and temporary. Once a data record is inactive, data that are classified as permanent must be archived and temporary data must be disposed of carefully.

Archiving is the process whereby data that have lost their relevance are moved to an alternative storage site for historical collections, an archive intended for the long-term retention of permanent data that are no longer meant for primary use. Data will no longer be maintained or updated, but merely serve as a historical reference. It is common for

archived data to have direct identifiers removed, depending on what further purposes legacy data can realistically serve.

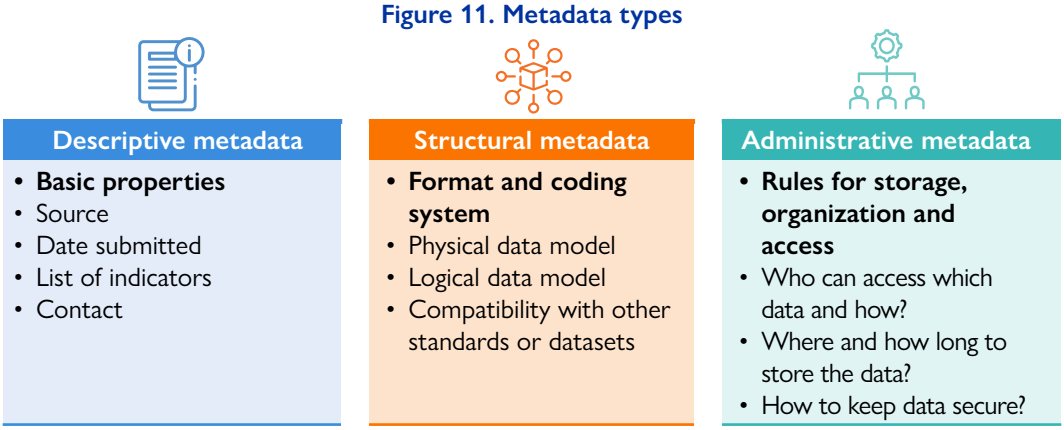
The determination to destroy data is based on similar factors. Understanding the legal environment and following the checklist of relevance are both steps in the right direction. The decision to destroy data nevertheless requires a bit more due diligence, as the action, if done properly, is permanent.

How to codify roles and rules in good documentation/Metadata

The rules and roles need to be codified formally in a series of documents or metadata.

Most people are familiar with the term "metadata" in its most commonly used form, that is, the documentation of data-set attributes, including collection process, time and source. But this is only one type of metadata (descriptive). The term metadata means *data about the data*. There are three main types of metadata – descriptive, structural and administrative, see Figure 11 – and each of them has a different job in the overall documentation of the data governance framework.

Figure 11. Metadata types



Descriptive metadata is documentation that records the basic properties of data sets. It is the basic information that usually accompanies an open-access data set when it is downloaded. The other two types of metadata, structural and administrative, are primarily used behind the scenes by the organization or agency and provide much of the core information on the data governance framework, or the well-defined roles and rules on data management. Structural metadata describe the data format and coding system needed to meet FAIR data standards and to be usefully combined with other sources of data, governmental or otherwise. Administrative metadata detail the rules for storage, organization and access. Annex 5 provides more details on each type of metadata.

SETTING UP A DATA GOVERNANCE FRAMEWORK IN AN INTER-AGENCY ENVIRONMENT

In contexts where data are provided by diverse organizations, different stakeholders may have different rights over the data and may have divergent preferences for what the data can and should be used for. When it comes to administrative data on trafficking in persons, the data may be very sensitive and data-sharing arrangements may involve multiple conditions and restrictions on downstream use of the data. The roles typically involved in decision-making will need to be carefully crafted to fit the situation.

Some of the main questions to consider in this type of inter-agency environment are the same as in other contexts but may be more challenging to address. For instance, who has rights to the data? Who determines and upholds these rights? Who decides for what purpose the data can be used and by whom? Other questions also arise: how is data stewardship organized and governed at the inter-agency level, if the aim is to bring together and use data assets from a range of organizations? What if multiple agencies are contributing to the same data asset, as in the case of a national referral mechanism? Is data stewardship vested only in one agency that has all the authority, or is it a shared responsibility, or can others have a say through an inclusive governance framework? How can the decision maker allow for the rights and preferences of relevant stakeholders? Do multiple stakeholders need to be informed and consulted before a decision is made? Will their formal agreement have to be sought and are they also decision makers? If so, are decisions made by consensus or majority? Can individual organizations opt out of certain processes? While there is no single best answer to these questions, they all must be answered in an inter-agency data governance framework.

While the roles underpinning data governance frameworks take a standard form in many institutional contexts, special attention has to be paid to how they need to be tailored to the kind of inter-agency environments that are often required for the management of data on trafficking in persons.

As explained in the section on "Roles" above, roles must exist to maintain and implement standards and facilitate decision-making and problem-solving related to the protection of data subjects, among other issues. In an inter-agency environment, there will be multiple data stewards, since each data-producing agency will be stewarding its own data assets. However, when data assets become shared among, or need to be managed in the interests of, a wider group of stakeholders than just one data-producing agency, the data stewardship role and functions may become shared by multiple agencies. Even if there is a primary lead for implementation, decision-making and accountability may become shared through an inter-agency data governance framework.

More broadly, Figure 12 provides an example of how to structure data governance responsibility in an inter-agency setting and illustrates one possible approach to assignment of responsibility and the chain of accountability. The steering committee is responsible for policy and high-level decisions, while the trustee is in charge of day-to-day decisions and implementation. In the diagram, the data stewards are those in the respective agencies.

Figure 12. Example of how to structure data governance responsibility

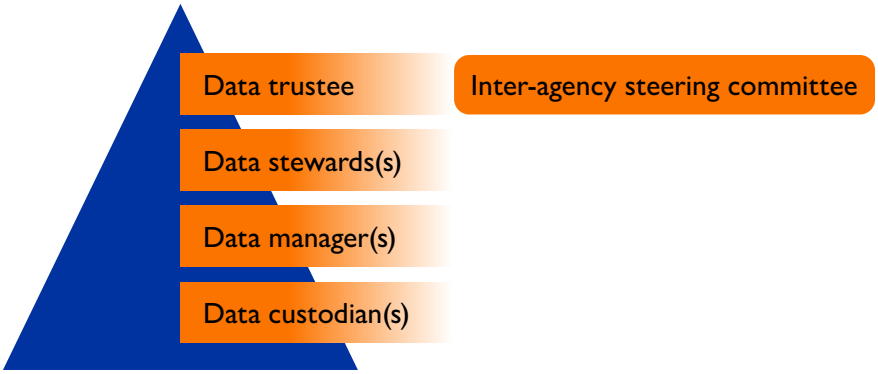


Figure 12 also indicates the role of an inter-agency steering committee, to create cohesion and consistency between what may be diverse data holders and sources. The committee members will play an essential role in developing data governance policies that satisfy the interests and needs of each collaborating agency. They have intimate knowledge of the data assets and operating procedures of their respective organizations, and will be able to advocate for those interests and needs while having a say in how data are governed and used.

Any team, no matter how it is composed or formed, will need specific and clear terms of reference that should be part of the data governance plan (or constitute a first step in the development of the data governance framework). This is especially important to meet the needs of all data stakeholders and to address any inter-agency conflict that may arise when the multiple entities involved have divergent interests. The terms of reference may be used by the data trustee or the steering committee to solve problems.

Examples of inter-agency arrangements

The previous chapter described the usefulness of the national referral mechanism model, designed to enhance coordination of case management and victim support, for facilitating better data collection. It also presented other types of strategy used by sourcing agencies and their strategies for collecting data from multiple stakeholders. The models are rediscussed here, with an eye to their function in improving the process of data governance for the sourcing and management of data.

National referral mechanism governance framework

Under an inter-agency data governance arrangement within a national referral mechanism, data may be primarily stewarded by a single, central coordinating agency that is responsible for processing and managing data sourced from other agencies. The agency may also serve as the central data repository and the official source of national statistics (if the specific arrangement allows for it to use the data in this way). The institutional framework developed to support and govern the national referral mechanism can be built on to govern the stewardship of data flowing through the system that supports it.

To date, few national referral mechanisms have fully developed data governance arrangements supporting this level of data management (beyond that needed for case management/victim support). The few that do exist, however, obviously have different approaches to different needs and challenges, particularly when it comes to how centralized decision-making authority and the role of data steward are.

For example, in the United Kingdom, the Home Office runs an information management system that supports the functioning of the national referral mechanism and that is used by other agencies as well. The Home Office essentially acts as the custodian and steward of the data from the point of creation and is vested with the rights to use the data for anti-trafficking evidence purposes, for reasons of substantial public interest. The Home Office further ensures that data stewardship is inclusive of other stakeholders through regular consultations and encourages comments or questions about the data through an email account linked to the quarterly publication of the statistics produced via the national referral mechanism.

Within the national referral mechanism of the Kingdom of the Netherlands, the NGO CoMensha acts as an independent clearing house, or data trust, between diverse

data-producing agencies and the government. Under this data governance model, while day-to-day stewardship of inter-agency data assets are the responsibility of CoMensha, decision-making is ultimately more decentralized, as the ownership and rights to data remain primarily with the original data-producing agency. The only data assets transferred or shared are those that stakeholders have agreed to.

Alternative inter-agency data governance structures

Chapter III also described how data sourcing can be streamlined through alternative institutional arrangements, including government agencies mandated to serve as national repositories and independent bodies that are set up outside the government to centralize decision-making authority and/or data stewardship in a completely or partially independent and autonomous manner. It is important to delineate the way these models may also be used (or replicated) to set up optimal data governance.

The previous section gave the example of CoMensha, an independent clearing house used where there is a national referral mechanism, but the same model of an independent clearing house could be applied in other contexts where there is no such mechanism.

Statistics Canada uses a more centralized model of data governance that may be more suitable for a national statistical office. Data-producing agencies submit data assets directly to the NSO, which hosts a data repository and plays the role of data steward for data coming from many government agencies, CSOs and other partners. The Statistics Act grants Statistics Canada the authority to obtain administrative data from organizations. This requires a great deal of trust, which Statistics Canada has worked to build and maintain over decades. Stakeholders, while no longer retaining rights over the data, continue to have strong advisory roles in the process and are informed about how the data will be used at every step of the way. The Statistics Act contains very detailed provisions on responsibility for the data, outlining chains of reporting and accountability, and processing instructions (what data, how and by whom). While stewardship is centralized at Statistics Canada, it is exercised within an extensive data governance framework with multiple independent checks and balances. These include prescribing an ethical framework and establishing an ethics committee, which must assess the NSO's justifications for using the administrative data. There are also multiple advisory committees supporting Statistics Canada's governing bodies when it comes to stewardship over these kinds of data, including policy, courts and correction advisory committees.

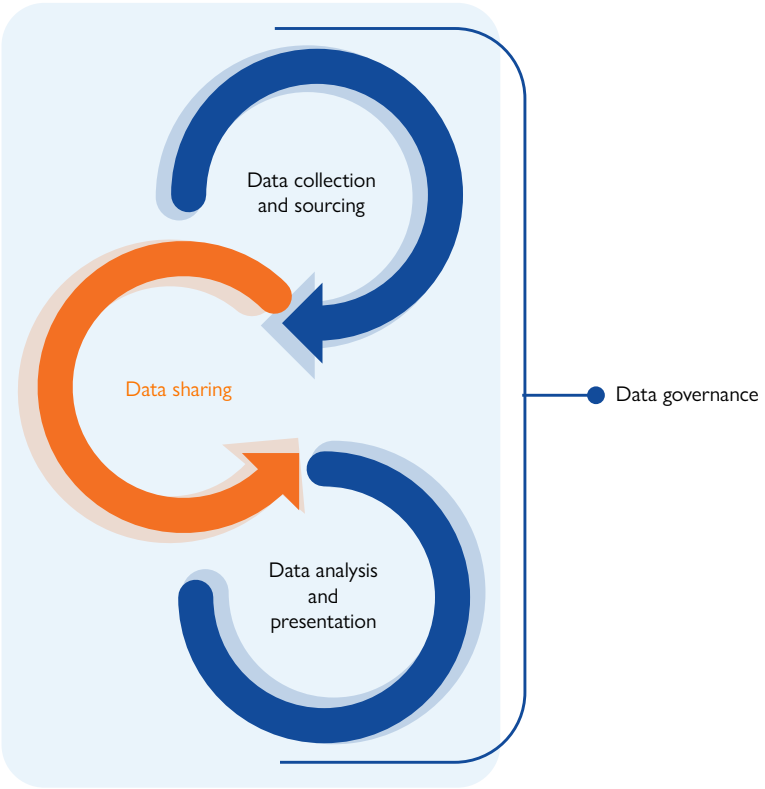
An alternative, or complementary, approach is to assemble an inter-agency task force. Establishing a task force involves bringing together personnel already employed by the relevant agencies to focus on TIP data. However, it is important to note that a task force is not a legal entity and therefore cannot be the custodian of any data assets. It may be a forum for coordination, discussion and decision-making, even if stewardship and custodianship remain with the data-producing agencies. New, shared data assets are not created but existing data assets may be better leveraged to serve the shared purposes of a wider group of stakeholders. Consultations have offered insights into the effectiveness of this strategy, which is relatively common but has had varying results. According to some familiar with the process, task forces are better at achieving data governance goals when they are more stable (formalized) and have clear, specific mandates. Task force members should also focus narrowly on TIP data management/collection rather than being assigned many other roles.

CHAPTER 5:

SHARING AND DE-IDENTIFYING ADMINISTRATIVE DATA

Data are only inherently helpful once they are used – for example, to detect trends and patterns, to determine traits that may be causal or related to other phenomena, or to assess the impact of a treatment or policy. While some data may be used (analysed, interpreted, communicated) internally or by the government agency that collects and manages them (practices that are discussed at length in Chapter VI), much of the data generated for a national repository or centralized agency will not be analysed in-house and instead will need to be prepared for sharing and use by others (see Figure 13).

Figure 13. The data life cycle: data sharing



In Chapter IV, the discussion centred on accountability and how it is assigned through data governance frameworks. Once data are shared externally, it is no longer possible for data stewards to protect the information or to apply the same rules to safeguard the privacy of data subjects outside the central agency holding the data. This does not free the agency of its responsibility vis-à-vis data subjects and providers. Instead, it becomes incumbent upon primary data holders to prepare data in such a way that they can be shared safely.

There are multiple ways to treat data so that they are safe to share and use, but each has implications for how useful the data are for particular purposes. The more details the data provide, the more useful they may be to university researchers, international organizations and anyone else interested in analysing them. There is nevertheless a trade-off between the level of detail offered and securing the data set from risk, as will be made clear below.

In this chapter, the discussion will focus on ways to ensure that data can be shared both privately and publicly while upholding the principles protecting data subjects, providers, users and stewards. It will also cover the secure sharing of data sets with external partners that have data-sharing agreements. Finally, the due diligence considerations discussed at the end of the chapter will allow government agencies to develop protocols for publishing data that have been carefully prepared for public use.

PRIVACY ISSUES WITH DATA SHARING AND PUBLISHING

Issues of data protection are just as pressing during the data-sharing stage as they are in earlier stages of the data life cycle. The data-sharing process has legal and ethical implications and poses numerous risks (see Box 6), for which safeguards must be managed and maintained. Doing so properly will be essential to upholding the legal and ethical standards of data protection.

Box 6. Privacy risks posed by data sharing

Individual privacy risk

Personal data, even after direct identifiers such as names or addresses have been removed, can pose a threat to individual privacy as unique attributes can be linked to identify a person: *this must be person X because no one else has these specific traits.* A breach of privacy, especially in the case of vulnerable people, is problematic whatever the source. In the case of highly sensitive data on victims of trafficking in persons, in the worst-case scenario the source may even be the trafficker, putting the victim in danger.

Cohort privacy risk

Even if there are not enough combinations of unique identifiers in the data to identify an individual, rare attribute combinations can be linked to known groups: *case records with specific attributes must be from these five individuals.* This is problematic as well, as someone could then identify a small, specific group of victims based on their knowledge of a form of exploitation, gender or nationality, posing a threat to the entire cohort. Publishing data may pose different risks for different types of cohorts in different contexts. Consider, for example, what might happen if data are published on ethnic groups in a specific location, especially if one of the ethnic groups is the target of violence.

Safety risk

There is also the possibility that an individual who cannot be directly identified can at least be assumed to be a member of an identifiable cohort, i.e. similar group/cohort sizes can indicate the likelihood that an individual is a member of the group.

It may seem that all of the effort involved, and the risk entailed, would preclude the sharing step of the data cycle altogether. This is a fairly common predicament for administrative and other types of sensitive personal data.

Once the data subjects and data-producing agencies have done the work needed to provide the data assets, it is important that those assets be used to provide better services and policy for all concerned. Data are only useful if used. But any use of data, particularly by third parties, comes with a potential risk: that confidential information will be exposed on data subjects. This is not only a breach of trust; it is also a violation of rights.

There are many options for processing data to remove identifiers (direct and indirect) so as to reduce the risks of violating the data subject's privacy when the data are shared. Unfortunately, there is a major trade-off between how useful data can be for analysis and to what extent the data are de-identified.

Trade-off between data utility and protection

The most useful data for research and analysis are generally to be found at the most granular level (such as the individual level) and could be used to (re)identify a data subject. Not only

do such data provide the most information to analyse and explore, they can also be used to track unique individuals or events across different administrative data sets or to detect repeat trafficking, especially if they include direct identifiers such as names or social security numbers. In some cases, they may also serve to link one data set to other data sources, to analyse the relationships between diverse phenomena.

However, using data with direct identifiers, whether individual names or an identification number, or even using de-identified data that still include personal traits represented as text (e.g. when a row of data has information on an individual's gender, birthplace, occupation or family size that can potentially be connected to identify them), introduces a risk to data subject privacy. At the other end of the spectrum, when data are transformed (or summarized) so that no individual traits remain, this presents less risk to data subjects, but also leaves less granularity for analysis. Table 3 lists basic forms of data and their characteristics.

Table 3. Basic forms of data and their characteristics

Type*	Definition
Raw data	Data collected on a person/event/organization and not processed In other words, raw data have not been modified (e.g. for the purposes of de-identification or aggregation).
Partially de-identified disaggregate data	Data collected on a person/event/organization and modified only marginally, by removing direct identifiers (see definition in the next section).
Aggregate/tabular data	Data combined and presented in a summarized format in the form of statistics, tallied counts, etc.
Synthetic data	Data that are artificially created rather than obtained by direct measurement (but preserve the statistical properties and relationships from the original data).

* This is not an exhaustive list of data types, as qualitative data in text format would be described differently.

Analysts usually use data sets in **raw form** to conduct a wide range of secondary quantitative analyses, including deciphering trends, making comparisons between groups and discovering relationships between factors that can inform the fight against trafficking. Raw data nevertheless hold information on data subjects that can be used to reveal who they are, even after individual identifying information has been removed (as in **partially de-identified data**): when a row in a data set contains multiple units of information about an individual case or victim (e.g. nationality, gender, industry, age), someone could potentially deduce a person's identity without their name being listed.

Data that are in **aggregate form** present tallies or percentages of cases in groups. This can be much safer than raw data, as by separating information into pockets (per cent of victims of a certain nationality or victims by gender), the risk that individual data subjects will be identified via their individual traits is reduced – but not eliminated, as will be demonstrated below. However, the availability of aggregate data alone severely limits the ability to use the data for further analysis (more details are provided later in this chapter).

Novel practices of data anonymization, such as the creation of **synthetic data**, have helped bridge the divide between making data useful to analysts and protecting the privacy of data subjects.

There are also several methods of de-identifying data (through pseudonymization or anonymization) that result in varying degrees of data protection, ranging from the simplest and least secure (removing name identifiers) to a much higher standard (e.g. achieving differential privacy). This chapter describes the different forms of data de-identification and their merits and drawbacks.

The following sections address the complexity of sharing data and cover concepts related to data privacy and risk. The discussion will then turn to the various ways in which data can be processed to protect confidentiality and the choices that must be made about what data to share (or not to share), who to share it with, how to process them and what to do with "real" data that cannot be openly shared.

WHAT ARE PERSONAL DATA?

How does one determine whether data are personal? It may seem like it would be a simple exercise of verifying whether or not the data list an individual's name, address or other information (see Table 4) that could make their identity directly known. While these very explicitly direct identifiers are certainly clear examples of personal data, they are not the only kinds of information that end up exposing an individual's identity.

Table 4. Examples of types of information clearly considered personal data

Direct identifiers
Name
Address
Email address
Passport number
IP address

Indeed, it may be possible to identify a person through a combination of information, such as age, place of birth and gender.[55] These are indirect identifiers. Both direct and indirect identifiers constitute personal data. In fact, the United States National Institute of Standards and Technology defines personal information[56] as:

> any information about an individual maintained by an agency, including:
>
> 1. [Direct identifiers:] any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and
> 2. [Indirect identifiers:] any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.[57]

[55] The European Commission defines indirect identification as follows: "Contrary to direct identification of a Statistical unit from its formal Identifiers, Indirect identification uses combined information elements to identify specific units. The Variables which are combined are called 'Indirect identifiers'. Examples of 'Indirect identifiers' are place of birth, race, religion, weight, activities, employment information, medical information, education information, and financial information." (Collaboration in Research and Methodology for Official Statistics (CROS) portal, accessed 8 September 2021).

[56] Known to many, and often referenced in the literature, as personally identifying information or PII.

[57] See https://csrc.nist.gov/glossary/term/personally_identifiable_information.

Article 4 of the European Union GDPR makes the same distinction:

> "personal data" means any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;[58]

In other words, information (whether a single trait or a series of traits linked together) that can be traced back to reveal an individual's identity remains personal data and poses the same risk to data subject privacy as an individual's name or other clear forms of identification. This means that the process of ensuring that data are also free of indirect identifiers before they are shared will often be much more complex than simply removing names and birthdates.

Table 5 gives an example of a fictional data set in which a string of characteristics can make it easy to identify the individual data subject. Even when classic unique identifiers like name and birthdate are obscured, the combination of highly specific traits, like exact age and birthplace, can lead to re-identification.

Table 5. Example combination of (rare) attributes

Name	Birthdate	Gender	Age	Nationality	City of birth	Country of residence	City of residence
-hidden-	-hidden-	F	34	Slovenian	Koper	France	Annecy

Even when the data are provided in aggregate instead of raw form, the risk of identification from combining traits, though reduced, may not be entirely removed. Individuals can still be identified in aggregate. Consider Table 6, which contains rare combinations of (rare) attributes that narrow down the number of individuals in a category to only one (providing a first look at why k-anonymization, which is described below, is important).

Table 6. Frequency of cases with a unique combination of traits

Trait 1: Gender	Trait 2: Nationality	Trait 3: Occupation	Trait 4: Age	Trait 5: Country of birth	Number of cases in data set
Female	-	-	-	-	534
Female	Canadian				178
Female	Canadian	Engineer			32
Female	Canadian	Engineer	48		4
Female	Canadian	Engineer	48	Nigeria	1

In addition, and as always in aggregations, the cohort privacy risk remains – consider, for instance, the risks of publishing the number of unaccompanied children at a displacement site.[59] Unaccompanied children are very vulnerable in most contexts and publicizing their presence at a particular location could lead to them being targeted by potential abusers. The

[58] See https://gdpr.eu/article-4-definitions/.

[59] The cohort privacy risk is defined in Box 6. Even if there are not enough combinations of unique identifiers in the data to identify an individual, rare attribute combinations can be linked to known groups.

same holds true for membership of a particular ethnic group; if published, such information may lead to individuals being targeted in an ethnic conflict.

If data containing personal information are so high risk and cannot be easily shared or published for public use, why use them in this form at all? Why not always aggregate at a level that makes it impossible to breach privacy? The answer is that this is hard, if not impossible, to achieve without severely limiting the usefulness of the data – we come back to the trade-off, introduced earlier in the chapter, between data utility and protection. Fortunately, even though highly sensitive data cannot be easily shared, there are novel workaround solutions, including the ability to analyse data without ever possessing them (this is addressed towards the end of this chapter).

Understanding how data may serve to identify data subjects, even after direct identifiers have been removed or the data aggregated, is the first step towards proper data-sharing risk assessment and mitigation. The next step is to begin the de-identification process.

WHAT ARE DE-IDENTIFIED DATA?

When are data sufficiently de-identified for a given purpose? What needs to be done to make the data shareable with external users and at what cost to the original data?

There are several methods of de-identifying a data set (or reporting the data characteristics in a way that direct or indirect personal information is not involved), and each has its own levels of privacy risk and analytic utility.

First, however, it is important to be clear about what is meant by de-identifying data by making it pseudonymous or anonymous.[60]

Pseudonymized versus anonymized data

According to Article 4(5) of the European Union GDPR:

> **pseudonymisation** means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

It is useful to note, as pointed out by Mourbey at al., that in the European Union GDPR definition, pseudonymized data are still personal data: "The data processed per Article 4(5) evidently still *relate* to an identifiable natural person; pseudonymisation merely prevents the *attribution* of the data to a natural person."[61] Mourbey at al. further add:

> The word "pseudonymisation" in the GDPR thus refers to a process which reduces the risk of direct identification, but which does not produce anonymous data. Pseudonymisation is referred to as a means of reducing risks to data subjects, and as an appropriate safeguard for any personal data used for scientific, historical or statistical research.

[60] There is some debate about whether data can ever be considered fully "anonymous" (with zero risk of exposing data subjects). The term is used here to describe a range of reduced risks, rather than an absolute. It involves a process by which re-identification is made difficult or impossible for both data holders and unauthorized third parties, whereas "pseudonymization" is designed to hide personal information in such a way that re-identification by data holders is still possible.

[61] Miranda Mourbey et al., Are "pseudonymised" data always personal data? Implications of the GDPR for administrative data research in the UK, *Computer Law & Security Review*, 34(2):222–233 (April 2018).

Anonymization, on the other hand, refers to the complete removal of all personally identifying information in such a way that it cannot be reverse engineered or otherwise restored. Full anonymization is more difficult to achieve and requires that even highly advanced re-identification techniques, such as those that a skilled attacker might use, could not lead back to the data subject's identity.

The type of processing that makes data pseudonymous or anonymized has implications for which third parties can use the data and for what purposes. There are many issues to consider.

Legal basis for processing and sharing the data

As explained in the section "Ensuring the legal basis for using data downstream is planned for in upstream primary data collection", there must be a legal basis for data use. For example, consent from data subjects for use of their personal data may be required for different kinds of, or any, purposes, including use by a third party. This may also apply to data derivatives, even if they are so heavily de-identified that they may no longer be considered "personal data".

Levels of risk

In the case of highly sensitive data (e.g. volunteered by highly vulnerable populations), sharing too much data can do more than merely violate the data subject's privacy – there may be a real risk of harm.

Levels of utility

As mentioned earlier, and related to the level of risk, there is a trade-off between removing potentially identifying, detailed information to protect privacy and providing highly granular data that can facilitate more robust analysis. It is necessary to evaluate what data can be safely shared with a given entity (taking into account consent, data governance frameworks and the usefulness of the analysis to be conducted), and to consider whether those data are still useful for the purpose to be accomplished.

The next section describes actual methods of de-identification and how each relates to these issues of consent, risk and utility.

METHODS OF DE-IDENTIFYING DATA

How can levels of privacy be measured? Essentially by measuring the probability that an attempt to re-identify, or reverse engineer, the de-identified data, will be successful. The de-identified data set must also be able to withstand attempts at identification through cross-referencing with other data sources. The protection must hold for *any* data point in the database.

Before delving into the mechanics of each method to de-identify data, it is useful to understand the broader analytical framework underlying the techniques.

General principles of de-identification

Differential privacy

One main approach to altering "real" data for the purposes of de-identification is that of differential privacy. Differential privacy is not a de-identification technique, but rather a mathematical property held by data when they have been de-identified to a point beyond re-identification.

Differential privacy is achieved when a data set produces similar outputs whether it contains any single individual's data or whether those data have been removed, making the marginal impact on a data subject's privacy the same whether their information is in the data set or not. In other words, there is no way to tell whether an individual's information exists within the data, no matter how unique their information is compared to other data points, because any results from analysing the data are roughly equivalent with or without that particular data point.

There are various methods for achieving differential privacy, but all are based on introducing some level of random "noise" to the results or statistics generated from the data. In statistical terms, noise is error, or unexplained variability, that typically has unknown origins (usually because the data-generating process is imprecise and unreliable). The higher the level of noise, the greater the privacy. Consider the basic example set out in Table 7.

Table 7. Example of a data set to which differential privacy can be applied

Victim citizenship	Counts
Afghanistan	7
Albania	9
Algeria	25

Let us assume that the organization holding these data would like to make this table publicly available in an interactive dashboard, alongside other tables with other victim information, but is (rightly) worried that the victims could be easily identified, thanks to the low count and the potential for cross-referencing with other tables. What it can do is publish a table in which a random number is added to the actual counts, that random number being drawn from a chosen probability distribution.

The method offers a probabilistic guarantee against membership inference attacks aiming to identify the presence of a single individual in the data. However, it does little to mask the presence of small groups matching detailed queries, which can be better dealt with by using a method of k-anonymization.

K-anonymity

Another approach to de-identifying data is k-anonymity. The process works by dividing the entire data set into different sets of observations. Membership of each set is defined by some unique combination of attribute values (e.g. adult or child) for selected attributes (or selected columns) (e.g. age) in the data set and there are as many sets as there are possible combinations of attribute values. Common combinations lead to sets with larger numbers of observations and rare combinations lead to sets with smaller numbers of observations.

Many sets may also be empty. The smaller the set, the higher the risk it poses for possible re-identification of data subjects. Observations that fall into smaller sets are therefore simply removed from the data set. The minimum number of observations that must be in a set for data to be public is *k*.

For example, at the *k*=10 level of k-anonymization, when nationality, age group and type of exploitation are selected as quasi-identifying attributes, an observation will be retained in the data set only if that observation has identical values for at least nine other observations (*k* minus 1 equals nine) for those chosen quasi-identifying attributes. Conversely, if there are fewer than 10 observations in the data set with those characteristics, those observations will be removed out of concern that an individual data subject could be discovered based on this information. Figure 14 provides an example of how k-anonymity can be achieved in this way, with *k* set to 3.

Figure 14. Example of k-anonymity

Birthdate	Gender	Postcode
13/01/1990	M	33710
25/10/1990	M	33410
18/08/1989	F	55810
14/02/1982	F	94320
30/11/1982	F	94870
20/12/1982	F	94580

Year of birth	Gender	Postcode
~~1990~~	M	~~33~~***
~~1990~~	M	~~33~~***
~~1989~~	F	~~55~~***
1982	F	94*
1982	F	94***
1982	F	94***

A policy decision must be made about the minimum number of observations a set must contain (the value of *k*) in order for it to be retained as part of the data set for further processing/analysis/sharing. Note that sets are usually built on selected attributes in the data set, so that combinations including other attributes may still be uniquely identifying. This is why the selection of the attributes on which k-anonymization will rest (and of course the exclusion of the others) can be challenging and involves carefully running through potential re-identification scenarios.

The various methods of protecting data through k-anonymization or differential privacy have their own benefits and drawbacks, which must be weighed before being adopted. Nevertheless, each paradigm of privacy offers many opportunities for processing data in a manner that can provide adequate safeguards for the use and sharing of highly sensitive data.

De-identification methods

(Simple) de-identified data

There are several ways to de-identify data. The simplest is to remove direct personal identifiers in data, such as names or addresses (see Table 8). Some indirect identifiers may also be removed, which will reduce risks but worsen utility – the difficulty lies in knowing which ones to keep and which ones to remove.

Table 8. Simple de-identification

Risk	Utility
High	High
Benefits	**Drawbacks**
Maintaining all other characteristics of the data set allows for the most complete analysis. This includes risk and protective factors,[a] as well as trends, especially when the data set consists of many indicators measured over time. Keeping all of the rows in the data set (e.g. all the victims, perpetrators) also serves to provide evidence on more marginal profiles and, if data are collected over time, to monitor their evolution.	The major risk of simple de-identification is that re-identification is relatively easy. Rare attribute combinations pose a threat to privacy regardless of whether the main identifier, like someone's name, is removed. The risk can be somewhat mitigated by reducing the detail available in the data set, for instance, replacing the exact age with an age bracket (e.g. 23 becomes 18–24), or top- and bottom-coding (e.g. not displaying ages above 50, or below 10). However, this cannot be done for all variables (e.g. gender) and may only marginally reduce the risk, depending on the nature of the data.

[a] Risk factors contribute to vulnerability, whereas protective factors improve capabilities to avoid, cope with or recover from harm (see IOM, *IOM Handbook on Protection and Assistance to Migrants Vulnerable to Violence, Exploitation and Abuse*, Part 1 – The determinants of migrant vulnerability (Geneva, 2019)).

Aggregating data

A data set is aggregated when only the descriptive characteristics of each indicator are reported, separately or in cross-tabulation with multiple indicators, such as the proportion of victims of trafficking in persons that are of a certain nationality (overall or disaggregated by gender or age), or percentages of traffickers that use a certain recruitment tactic (overall or by region).

A related solution is to publish aggregated data through interactive dashboards – this means users are better able to "interact" with the data. For instance, users might be able to filter the rest of the charts by gender or majority status. The risks are the same as for aggregate data in general – namely, that it must not be possible to narrow the data down to a small group of easily identifiable individuals (see Table 9 for a description of the risks and utility of aggregating data). This will likely be more challenging to implement for an interactive dashboard than for traditional "static" data. An additional risk may arise, depending on the modalities of publication of the dashboard and the platform used. Care must be taken to ensure that the raw data generating the statistics cannot be accessed through the dashboard, and therefore it may be necessary to pre-compute all the possible aggregate counts and "cross-counts" to populate the dashboard.[62]

[62] In theory, such pre-computed counts and cross-counts could also be published as they are, and not as an interactive dashboard. However, the resulting data set would be large, unwieldy and not suitable for exploration using conventional tools or methods.

Table 9. Aggregating data: risk and utility, benefits and drawbacks

Risk	Utility
Low-medium (when data are aggregated above a particular threshold of cases – see Table 6)	Low
Benefits	**Drawbacks**
Ability to report basic statistics on the issue Removes risk (mostly) by pooling individual data points into primary categories, making it difficult to identify individual data points	The risk is still not zero. Releasing small counts of individuals, or precise counts/statistics that may be combined to produce small counts, may still be sufficient to link individuals to aggregated data sets (see, for instance, Table 6). The aforementioned cohort privacy risk, whereby rare attribute combinations can be linked to known groups, also remains. Mitigating these risks by identifying possible privacy leaks may be an ad hoc and labour-intensive process. In terms of analysis, the problem is that when information is siloed into discrete containers, researchers can no longer make connections between traits to determine if there are relationships between gender and a particular type of exploitation, for instance. Presenting the data clustered by cohort leaves room for some analysis of patterns but not for complex statistical analysis.

K-anonymization

The concept of k-anonymity is explained earlier in the chapter. In short, k-anonymization consists in removing observations with rare combinations of attribute values that fall in sets of fewer than *k* individuals. Observations with more common combinations of attribute values remain in the data set. Table 10 describes the risks and utility of k-anonymization.

K-anonymization shares some similarities with data aggregation, in that rare profiles are (or should be) redacted for the individuals' safety. Depending on the data set, k-anonymization may still allow for more multivariate analysis (such as regression analysis, for instance).

Table 10. K-anonymization: risks and utility

Risk	Utility
Low to high (depending on the value of *k* and on the data set itself)	Medium
Benefits	**Drawbacks**
The benefits are similar to those of simple de-identification, except that marginal profiles can no longer be studied (in turn, they also cannot be matched with rare profiles in the real world). It nevertheless remains possible to analyse the main general trends by keeping more common profiles.	Re-identification may be more difficult to accomplish with k-anonymization, especially when *k* is set to a higher number. However, while the risk to individual privacy is mitigated, there may still be a risk that a unique cohort can be discovered. The redaction of outliers can also lead to significant data loss, depending on the data set and the chosen value of *k*.

Creating synthetic data

Generating synthetic data is a process by which a new data set is synthesized in which the records do not correspond to actual individuals, but which preserves the structure and statistics of the original data. It involves altering individual data points or responses in such a way that the data's general structure is preserved but enough uncertainty or random noise is introduced to thwart anyone trying to re-identify data subjects. In other words, synthetic data capture the structural and statistical properties of a sensitive data set, although there is not an exact 1:1 correspondence between the records of the sensitive and synthetic data sets. The technique can also suppress any attribute values that are rare in the data set (e.g. a case of trafficking for blood, organs or body parts) and combinations of rare attribute values (e.g. citizenship from a small country, together with place and type of exploitation), achieving k-anonymity.

Table 11 describes the risks and utility of synthetic data. A synthetic data algorithm developed by IOM and Microsoft is described in Box 7 and Annex 6.

This type of treatment of sensitive government data that need to be shared for research purposes is becoming more popular. For example, the United States Government creates synthetic data for public use when processing census data that are to be shared.[63]

Table 11. Synthetic data: risks and utility

Risk	Utility
Very low	High
Benefits	**Drawbacks**
This technique allows an unredacted data set to be shared while preserving the victims' confidentiality.[a] As such, its benefits are similar to those of simple de-identified data, without the resulting risks – no minor consideration. In particular, the data remain useful for analysis (unlike with aggregation), while data subjects are protected from re-identification.	Synthetic data can be technically difficult to produce.[b] Synthetic data can only preserve both privacy and utility when data dimensionality is low, i.e. there is a great deal of overlap between records. This is explained in greater detail in Box 7 and Annex 6, but in short, and similar to k-anonymization, the method will not work well on data sets with data subjects that are all very different from each other. Analysis of synthetic data will result in generalizations about the data properties of the original data that are not completely accurate. If the record corresponding to an individual is not reproduced in the synthetic data set, then the individual cannot be re-identified (since the record is not available for linking). However, if distinctive combinations of attributes relating to an individual are present in the synthetic record, then, depending on how much is known about how the synthetic data set was produced, this

[63] The United States Census Bureau, which produces and shares synthetic data with researchers for analysis, has an agreement that allows researchers to submit their analyses for validation by government officials with access to the real census data. While it may not always be practical for a national repository or government department to offer this service, it is an option that can be mutually beneficial to governments and research partners. See www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html.

Benefits	Drawbacks
	may or may not suggest that the individual was in the original sensitive data set. The method used to synthesize data may or may not control this risk of "membership inference". More broadly, drawbacks depend on the level of sophistication of the data-generating algorithm. Some methods of creating artificial noise can make bivariate analysis less accurate. Finally, communicating the results of synthetic data analysis to a less technically savvy audience can be difficult. Issues of trust may arise when presenting analyses of data that are not "real".

[a] It is unredacted in the sense that it preserves all (or nearly all) attributes, even those that are rare in the data set.
[b] Resources are available online, notably on Microsoft's GitHub: https://github.com/microsoft/synthetic-data-showcase.

Box 7. The Microsoft/IOM synthetic data algorithm

IOM and Microsoft have collaborated to create a synthetic data de-identification solution through the Tech Against Trafficking Accelerator.[a] They used the CTDC data set to create and fine-tune the algorithm created.[b] Microsoft has made the algorithm freely available on GitHub,[c] and IOM and Microsoft jointly released the first CTDC synthetic data set in September 2021, together with interactive dashboards used to explore the data.[d]

The algorithm works by generating a data set composed entirely of attribute combinations that are common in the sensitive, original data set (common being defined as appearing at least *k* times). In the case of the CTDC, attribute combinations are combinations of characteristics of a victim of trafficking. A concrete, fictional example would be "Female, over 55, forced labour, exploited in the United Kingdom".

The algorithm preserves not only the structure of the data set but, more importantly, the statistical relationships between attributes.

More details, including a step-by-step description of how the algorithm works, are available in Annex 6.

[a] The paper describing the approach is available from https://arxiv.org/abs/2005.05688.
[b] See www.ctdatacollaborative.org/.
[c] See footnote 36.
[d] See www.ctdatacollaborative.org/global-synthetic-dataset.

DUE DILIGENCE CONSIDERATIONS FOR SHARING OR PUBLISHING DATA

Securely sharing data, whether internally or externally, requires decisions early in the data management process. Rules will need to be developed for sharing data internally or externally in light of the different types of data available and different types of users. Data stewards can rely on the risk classification system established in the process of creating administrative metadata (see Chapter IV) to determine what level of access can be provided to which internal or external data user. Several common principles nevertheless apply to all types of data sharing with all target groups, as set out below.[64]

[64] Much of the guidance in this section is aligned with the *IOM Data Protection Manual* (see footnote 26), which also provides useful considerations and templates.

Legal basis for processing, sharing or publishing data

Just as central agencies wishing to use administrative data produced by others for evidence purposes must have a legal basis to do so (see Chapter III), any sharing or publishing of data must be legally grounded. The legal basis is very much tied to the purpose of data collection, which is another reason why this purpose must be appropriately limited (see "Purpose" below).

Consent is one such legal basis, but by no means the only one. In fact, consent to share data is not applicable to all trafficking data in all contexts. In particular, it depends on the type of data –criminal justice data may not be published on the same legal basis as the characteristics of victims of trafficking assisted by counter-trafficking agencies. It may also depend on the level of de-identification of the data. A data set that is considered to meet the standards of anonymization (e.g. thanks to aggregation, synthetic data generation or k-anonymization with a large enough *k*) may no longer be considered personal data and may require a different legal basis for publication.

The *IOM Data Protection Manual* recommends that the issue be considered early, before the data are collected, processed and obtained – hence the need to identify potential third parties before data collection and, in the context of administrative data especially, to disclose all processes in the data life cycle to the data subject when asking for consent. In particular, data subjects must be told why specifically the data are being collected and shared – otherwise, their consent is not "informed".

Purpose

While data that are published for public consumption need to be available for any purpose, a request for transfer of personal or sensitive data must be clear and specific on the purpose for which the data will be used. It should include a description of the nature and categories of personal data needed and the method of transfer to be used. All disclosures to third parties should be on a "need-to-know" basis and only those categories of personal data needed to meet the purpose of the transfer request should be revealed.

These elements of the data-sharing process (purpose, data to be shared, method of transfer) should be formalized in a legal agreement, law or legal mandate, which should also stipulate the obligations of the data recipient with respect to the data (where to store them, when and how to delete them).

Proportionality

Before deciding whether to share or publish the data, the risks (to all involved) must be weighed against the benefits. Actual and potential risks should be identified, together with their likelihood and the resulting harm. Of course, data subjects are at the centre of these considerations, but the staff or organization collecting or managing the data should also be considered. Will data sharing breach trust? Will it cause reputational harm?

Here are some questions that can guide such an analysis.

- What is the nature of the data that are requested? Can individuals be easily re-identified from these data, including by cross-referencing with other data sources? Is there any way to share anonymized data instead?
- How long will the data need to be held?
- What entity is requesting the transfer and what is its relationship with the organization holding the data? Does it have the resources and knowledge needed to securely store and process the data? Could/would it cross-reference them with other data sets to achieve identification?

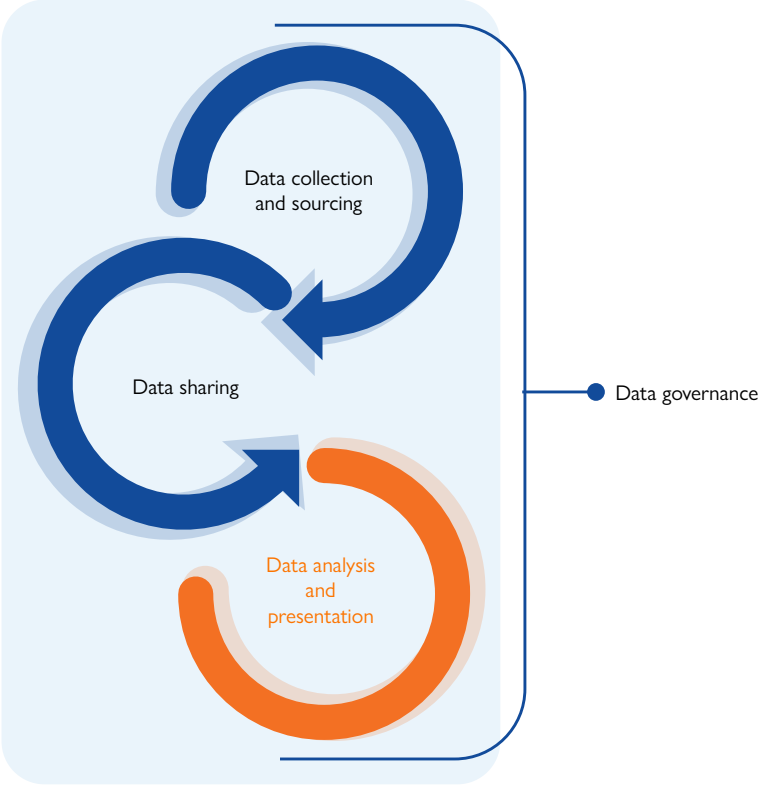
What will the sharing of data achieve? Will it help to learn more about trafficking and inform the response? Will it help prevent trafficking or improve assistance for the victims? Will it help to stop traffickers?

CHAPTER 6:

ADMINISTRATIVE DATA ANALYSIS AND PRESENTATION

The data life cycle culminates in its final phases of analysis, interpretation and presentation (see Figure 15 below).

Figure 15. The data life cycle: data analysis and presentation



The steps required to prepare various data types for use by external parties are described in the previous chapter. This chapter discusses strategies for using data for reporting purposes, including for the analysis, presentation and communication of administrative data.

Analysis of administrative data can provide a range of insights and contribute greatly to evidence-based action against trafficking in persons. Administrative data, by itself, can hold the information needed for many of the objectives listed in Chapter III. They can even be used to produce estimates of victims who are likely missed by other data sources and analyses, using techniques such as multiple systems estimation. The analysis and presentation of statistics and data visualizations can take many forms depending on the purpose of the analysis and the target group. While properly presented data can have a powerful impact in terms of positive action to combat trafficking in persons, improperly presented data can also be misleading and even damaging.

After data are released, their responsible use, interpretation and communication are no longer exclusively governed by the government agency from which they originated. It is therefore important, when making data available to the public or third parties, to offer guidance on responsible use and communication.

This chapter discusses how to represent administrative data accurately and avoid mistakes that can lead to misinterpretation. It starts by providing an overview of the TIP data evidence landscape, going into the value of administrative data, its strengths and limitations. It also highlights other types of data than can help inform the knowledge base on trafficking, including by complementing administrative data. The final section of the chapter provides concrete tips, describes good practices and presents the pitfalls to avoid when presenting data.

OVERVIEW OF THE TRAFFICKING IN PERSONS (AND RELATED) DATA EVIDENCE LANDSCAPE

The first, key step is to understand what administrative data can and cannot tell us about trafficking in persons in a given context, how to avoid misrepresentation, and how this type of data fits into the broader evidence landscape.

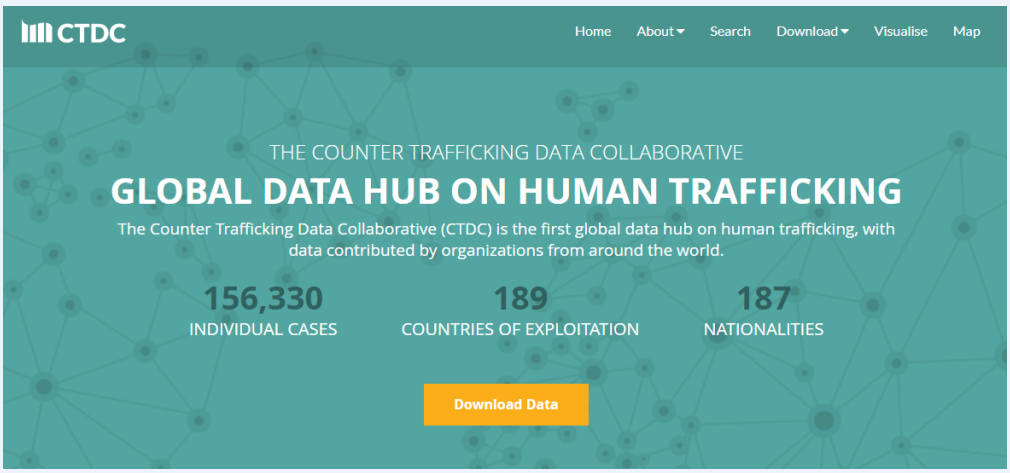
The aim is to identify what each type of data can offer for statistical analysis, as each can helpfully supplement the other. It is also beneficial to recognize opportunities to use other data sources in combination with administrative data, to make research findings more robust.

The use of administrative data for statistical analysis

As mentioned before, the analysis of administrative data can play a powerful role in enhancing understanding of social phenomena and discovering patterns and relationships. This is due in no small part to the fact that these data constitute one of the main, and often one of the only, windows onto the crime. They often provide detailed insight into the profiles and experiences of populations that are hard to reach using traditional survey methods (victims and perpetrators) and into the different forms of trafficking in persons (which is, by definition, a hidden crime). Hence the need to develop the capacity to collect administrative data and the related data systems. However, it is equally important to be aware of the characteristics and limitations of administrative data, to avoid misrepresenting results and potentially misallocating resources. Box 8 and Box 9 provide examples of how administrative data may be leveraged, and the insights that can be gained from their analysis.

It is becoming more and more routine to use administrative data for analysis purposes, especially as the capacity for this form of data collection expands. In fact, in some domains, such as education policy and social work research, the use of administrative data has become almost commonplace.[65] This is fairly intuitive, as governments were collecting school and social work case records long before the current data revolution. As administrative data in other areas become more accessible, so will their use in analysis.

Box 8. The Counter Trafficking Data Collaborative



[65] See, for instance, https://direct.mit.edu/edfp/article/12/2/129/10264/The-Promise-of-Administrative-Data-in-Education, https://pubmed.ncbi.nlm.nih.gov/25711312/ and https://repository.upenn.edu/cgi/viewcontent.cgi?article=1057&context=psc_publications.

The IOM Counter Trafficking Data Collaborative (CTDC) is the first global data portal on trafficking in persons, with disaggregate, primary data contributed by multiple agencies in a standardized format.[a] An unprecedented achievement in the field of migration data, the CTDC currently combines some of the largest TIP case data sets in the world, resulting in one centralized data set with information on over 156,330 cases involving 187 nationalities in 189 countries across the five regions (Africa, Americas, Asia, Europe and Oceania).[b]

The goal of the CTDC is to break down information-sharing barriers and equip the counter-trafficking community with up-to-date, reliable data on trafficking in persons. Facilitating an unparalleled level of access, the CTDC provides analysts, academics, practitioners and policymakers with the information they need to support effective counter-trafficking policies and programmes. This includes thematic data stories,[c] maps,[d] interactive dashboards[e] and de-identified data sets[f] for download.

In September 2021, IOM released the first ever synthetic data set of individual survivors of trafficking, in partnership with Microsoft. The synthetic data set is the largest collection of disaggregate, primary data on individual victims ever made available to the public; it has strong privacy guarantees that preserve the anonymity and safety of victims and survivors. The use of synthetic data enabled the CTDC to publish visualizations and dashboards that are far more interactive, since there will be no risk of an individual being identified. More details on the synthetic data algorithm are available in Chapter V and Annex 6.

[a] The CTDC is available from www.ctdatacollaborative.org/.
[b] The CTDC's partners are IOM, Polaris, Liberty Shared, the Portuguese Observatory on Trafficking in Human Beings (the CTDC's first government partner) and the NGO A21.
[c] Available at www.ctdatacollaborative.org/visualisations.
[d] Available at www.ctdatacollaborative.org/maps.
[e] Available at www.ctdatacollaborative.org/visualisations/interactive-data-dashboards.
[f] Available at www.ctdatacollaborative.org/global-k-anonymized-dataset and www.ctdatacollaborative.org/global-synthetic-dataset.

Box 9. UNODC's Global Report on Trafficking in Persons

Pursuant to the 2010 United Nations Plan of Action to Combat Trafficking in Persons, UNODC produces a biennial Global Report on Trafficking in Persons. The report draws primarily on official national information collected from countries all over the world. It presents data and analyses of trafficking in persons at the national, regional and international levels in a balanced, reliable and comprehensive manner.

UNODC is also advancing the work on testing methodologies for estimating the total number of victims of trafficking in persons. Not only will this help uncover the real magnitude of the phenomenon – which has so far been an elusive target for the trafficking research community – it will also assist countries in measuring progress towards the attainment of three Sustainable Development Goals (SDGs 5, 8 and 16 all set targets related to trafficking in persons).

Strengths of administrative data

As shown in Table 12, administrative data have many advantages that are useful for multiple research endeavours. In short, administrative data may offer large sample sizes on a population that is hard to reach, in a cost-effective way, given that these data have to be collected as part of an organization's operations. These data may also be collected consistently by different agencies and over time, enabling the study of trends and policy impact. Finally, administrative data may not suffer as much from certain types of bias as other forms of data, including survey data (for instance, recall and social desirability).

Table 12. Properties of administrative data and consequent benefits for analysis

Properties	Description	Benefits for analysis
Large sample size	Trafficking in persons remains a rare event, in the sense that many (perhaps several thousand) individuals would have to be surveyed before finding one who is involved in trafficking in persons. Relatedly, the populations of victims and perpetrators are of course hard to reach, also given the hidden nature of the crime. Therefore, administrative case data are often the only source of data available with a larger number of observations on victims and perpetrators; or at least the only source of data that are systematically collected over time, especially as governments are advancing efforts to improve administrative data collection, management and use.[a]	More data on target, hard-to-reach populations allows for better descriptive analysis. In particular, a high level of detail on each record combined with a large sample size means that the data could be used to identify patterns, trends, profiles, and the typologies of the populations being identified. This in turn can inform government responses.
Cost-effectiveness	Record-keeping is already a function of front-line agencies' operational activities (whether it is necessary for service provision, donor reporting, etc.), making the collection of administrative data more cost-effective.	Data can be obtained in timely fashion and sustainably.
Consistency	If proper protocols for consistent data collection are in place (see Chapter III) and there is capacity, regular standardized data will be available from stable sources.	Data collected for standardized indicators that are consistent over space and time can provide reliable information on developing trends. However, longitudinal analysis is useful not only to track the shifting nature of the crime/ trends over time, it is also particularly fruitful for the assessment of social policy[b] and of the effectiveness of service provision, which can then contribute to the development of improved social policies and programming.[c]

Properties	Description	Benefits for analysis
Reduces some forms of bias (social desirability and recall)	Information provided by a victim receiving services from trained front-line agency staff may be less prone to some forms of bias, including social desirability (reluctance to offer certain kinds of information for fear of being perceived negatively by the individual collecting the data) or recall (limited memory of events that may have occurred too far in the past).	Reducing these forms of bias improves data validity.

a D. Card, R. Chetty, M.S. Feldstein and E. Saez, Expanding Access to Administrative Data for Research in the United States, paper written for the National Science Foundation 10-069 call for white papers on "Future Research in the Social, Behavioral & Economic Sciences" (2010).

b Andrew Penner and Kenneth Dodge, Using Administrative Data for Social Science and Policy, *The Russell Sage Foundation Journal of the Social Sciences*, 5(2):1–18 (March 2019).

c Roxanne Connelly, Christopher Playford, Vernon Gayle and Chris Dibben, The role of administrative data in the big data revolution in social science research, *Social Science Research*, 59(09):1–12 (September 2016).

Limitations of administrative data

Table 13 provides an overview of the two main limitations of administrative data: coverage (or lack thereof) and the inherent focus on identified victims.

Table 13. Properties of administrative data and consequent limitations for analysis

Properties	Description	Drawback for analysis
TIP administrative data are only available where front-line agencies and other data-producing organizations are operational and able to collect and share such data.	Administrative data may not be available for all countries or districts of a country, and where data do exist, they may not always be comprehensive in terms of coverage (i.e. not all data-producing agencies may be able to collect and share data).	In terms of interpretation of data, this implies that large quantities of data on identified victims of trafficking do not necessarily indicate higher prevalence of trafficking in persons. Indeed, they may equally be indicative of an effective counter-trafficking response, and/or front-line agencies with good data-collection capacity.
TIP administrative data come from identified individuals or events (like most sources of data on trafficking in persons globally).	Identified cases are better understood as a sample of the unidentified population of victims or perpetrators, yielding insight into trafficking trends and patterns. A sample of identified victims or perpetrators may be biased if some types of trafficking case are more likely to be identified (or referred) than others.	The fact that the sample is not a random sample of the population of victims means that the assumptions of many tests of statistical significance are not met. Additionally, the extent of possible sampling bias is not always known nor able to be corrected for, since the unidentified population is, by definition, unknown. This is particularly the case if some of the counter-trafficking front-line agencies in a region focus on a specific type of trafficking (e.g. trafficking for the purpose of sexual exploitation, trafficking of children, of migrants, etc.), although in this case, the bias may be more easily identified and reported.

Given these two main limitations, it is also important to stress that the quantity of identified cases recorded by administrative data is not indicative of prevalence. Prevalence can be loosely defined as how much trafficking is happening in a given place at a given time. Of course, identified cases of trafficking can help shed some light on the types of trafficking that are more common, but they cannot (on their own) provide a definite picture of prevalence (see the last section of the chapter for more details). Nevertheless, there are some methodologies that use administrative data to estimate prevalence and to investigate never-identified profiles, as explained in Box 10.

Where administrative data come up short in terms of research on prevalence, more traditional forms of data may be able to fill the gap (see the section "Other data types (survey, geospatial, big data, narratives/qualitative)"). Furthermore, some prevalence estimates based on survey data also use administrative data to fill data gaps that traditional sampling methods cannot fill. These two sources of data can therefore complement each other – this is the case, for example, of the report *Global Estimates of Modern Slavery*,[66] which uses national probabilistic surveys and IOM's database (see Box 11).

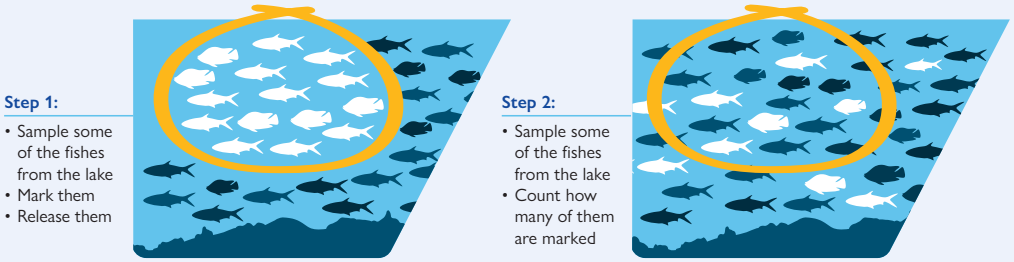
In any case, despite their limitations, these data are indispensable: they are often the only window of information onto the crime and provide detailed insight into the profiles and experiences of the victims, the forms of trafficking in persons, and information on perpetrators.

Box 10. Estimating prevalence

To estimate prevalence, or the proportion of a total population sharing some trait or affected by a phenomenon, a sample of the target population is needed that can be expected to be representative of the population as a whole. To assume that the sample is representative, it must be obtained by some form of random sampling. As administrative data are not obtained from random sampling, there is no easy way to tell whether they are representative of the larger population of trafficking victims. In addition, traditional survey methods that use probability-based sampling techniques – which are not used to obtain administrative data – tend to fall short in the context of largely hidden populations.

One of the techniques currently being used to estimate prevalence by means of administrative data is that of Multiple Systems Estimations (MSE). MSE takes existing lists of victims of trafficking collected by different organizations (for example, law enforcement agencies or CSOs) to estimate the number of victims who are not on any lists using a statistical methodology based on the concept of capture-recapture (see Figure 16). The conditions that must be satisfied to conduct an MSE in a given context are described in Annex 7.

Figure 16. Exemplifying the concept of capture-recapture: counting a lake's fish population



[66] ILO, IOM and Walk Free Foundation, *Global Estimates of Modern Slavery: Forced labour and forced marriage* (Geneva, 2022).

At the time of writing, MSE exercises to determine the prevalence of trafficking[a] had been conducted in eight countries and one capital city: Australia, Ireland, the Kingdom of the Netherlands, Romania, Serbia, Slovakia, the United Kingdom, the United States of America and the city of Madrid in Spain.[b]

MSE has nevertheless been used much more widely globally to estimate populations of sex workers, irregular migrants and people living with HIV/AIDS, to name but a few.[c]

[a] For general comments on using MSE in this context, see D. Durgana, and J.J.M. van Dijk, Measuring the Hidden Numbers of Human Trafficking Through Multiple Systems Estimation: Lessons Learned and Challenges Outstanding, *Journal of Crime and Delinquency*, 67(13–14):2188–2212 (January 2021).

[b] See Bernard Silverman, Modern Slavery: an application of Multiple Systems Estimation, November 2014. UNODC research briefs on the use of MSE to monitor SDG target 16.2 in the countries listed are available from www.unodc.org.

[c] Sheila M. Bird and Ruth King, Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy, *Annual Review of Statistics and its Application*, 5:95–118 (December 2017).
Jalal Poorolajal, Younes Mohammadi and Farzad Farzinara, Using the capture-recapture method to estimate the human immunodeficiency virus-positive population, *Epidemiol Health*, 39 (October 2017).

As capacity for data collection and management grows, so can the potential to collect a wider variety of data that can be used for multiple purposes to combat trafficking in persons. It is also important to define what each type of data can offer for the purposes of statistical analysis, as each can helpfully supplement the other.

Other data types (survey, geospatial, big data, narratives/qualitative)

To fully understand how administrative data fit into the overall evidence base on trafficking in persons, they must be considered in the context of other sources of data, including their advantages and limitations. This will also enhance understanding of how to combine diverse data types so as to shed light on relationships, drivers and what may be "working" (or not working) in terms of policy and programming (see the next section and Box 11 for an example of how two different data types can be combined).

Survey data

High-quality survey (or census) data with robust sample sizes that are collected from respondents through probabilistic random sampling methods are typically considered the gold standard of social science data, at least in terms of what they can offer for statistical analysis. When it is reasonably certain that respondents from populations of interest have a relatively equal probability of being selected with a survey instrument, and the instrument has been thoroughly tested with respect to valid, reliable indicators, limiting the various sources of potential bias, these data can be used to make inferences about the characteristics of a broader population.

As mentioned above, one of the problems with using survey methods to collect data on victims of trafficking in persons, or any hard-to-reach population, is that such traditional sampling techniques are often unsuitable for effectively sampling (relatively rare and hidden) target groups. The inability to reach these groups creates problems for inference as it cannot be assumed that respondents in TIP situations have an equal chance of being sampled and the survey is unlikely to produce a sufficient sample size for analysis. In addition, direct, probing questions to investigate trafficking will likely be challenging to ask due to ethical, safety and sensitivity issues. For example, surveying children directly can give rise to ethical issues and other challenges. This means that it is often the case that an adult primary respondent provides answers on behalf of children; however, evidence suggests that adults

tend to underreport the experiences of children.[67] In the aforementioned *Global Estimates of Modern Slavery*, this is precisely one of the reasons why IOM data were needed: the survey data did not provide sufficient information on children. Another example is sexual exploitation, which is challenging to ask about in household surveys in many contexts owing to the sensitive nature of the topic.

Referring to forced labour and trafficking in persons, the ILO usefully summarizes the difficulties with using traditional methods as follows: "These are difficult phenomena to survey for a variety of reasons: they are secret, criminal activities, the concepts are not self-explanatory and the people concerned may be unable or unwilling to acknowledge their situation and to identify themselves as victims."[68]

However, just as administrative data sources have improved in availability and quality, offering more possibilities for analysis, survey methods have also evolved to overcome some of the difficulties of sampling hidden populations. Many novel approaches to sampling rare and hidden populations are starting to be used for research on trafficking in persons. In a recently published introductory overview of prevalence estimation in modern slavery, the Global Fund to End Modern Slavery describes these survey methodologies as tailored to be most effective for hard-to-reach populations.[69] The overview provides a succinct, yet comprehensive, summary of survey and sampling methods such as respondent-driven sampling, the network scale-up method and multiple systems estimation.[70]

Still, it is important to note that, although more useful guidance has been produced in recent years, there are still no international standards or guidance on these matters.

Geospatial data

In the study of trafficking in persons, especially as it often relates to movement and migration, connecting administrative data to geospatial data for the purposes of mapping routes and corridors of trafficking in persons can be game-changing for police and other departments that must identify or predict locations where enforcement or assistance is needed.

Beyond mapping movements, other types of geospatial data, such as satellite imagery or geolocation data from mobile phones, can provide even more context and layers of information. Recent research on modern slavery has illustrated how layers of relevant geospatial information can detect relationships between climate-based degradation and exploitative labour practices or identify sites of possible labour violations and abuse in remote areas, to assist workers in agriculture, mining, fishing and other sectors.[71]

While there are many benefits to pulling in this form of data for analysis, there are also potential drawbacks. Most importantly, using this type of data may involve security risks, especially when visualizing locations at a more granular level (see the section "Assessing the confidentiality risks to avoid undermining data subject anonymity" below).

---

67 Eva Dziadula and Danice Brown Guzmán, Sweeping It under the Rug: Household Chores and Misreporting of Child Labor, *Economics Bulletin*, 40(2):901–905 (April 2020).

68 ILO, *Hard to see, harder to count. Survey guidelines to estimate forced labour of adults and children* (Geneva, 2012), p. 8. That being said, some forms of exploitation, such as trafficking for forced labour, seem to be more amenable to sampling using specially designed sampling methods.

69 Laura Gauer Bermudez, David Okech and Mihir Prakash (eds), *Methods of Prevalence Estimation in Modern Slavery. An Introductory Overview*, Prevalence Estimation: Methods Brief (Global Fund to End Modern Slavery, 2021).

70 It also provides citations and references to original methodology literature, for researchers interested in exploring these techniques further.

71 See, for instance, the work of the Rights Lab at the University of Nottingham, at www.nottingham.ac.uk/research/beacons-of-excellence/rights-lab/programmes/data/index.aspx.

Big data (machine data)

Machine data has confirmed potential to enhance detection of and offer new insights into the phenomenon of trafficking in persons. A common example is the mining of data from the dark web to identify victims and/or catch abusers, mostly in cases of sexual exploitation. Another example is web crawling to identify ads involving exploitative practices. Even as criminals adapt to evade detection, developing codes and symbols that can fly under the radar of the algorithms designed to find them, the technology is developing even faster through machine-learning techniques. Machine data, in the form of financial information, is also mined to detect anomalies in transaction histories that can indicate activity surrounding cases of trafficking in persons.

Qualitative data (narratives, open text)[72]

Qualitative data sources provided crucial insights into the phenomenon of trafficking in persons long before the need for prevalence estimates moved the field of study toward quantitative data collection. A great deal of important field research focuses on community-based case studies or the analysis of narratives. Typically, however, and by design, insights from studies that offer such deep, rich descriptions tend to be highly context specific and have not been connected to a broader national or international picture of trafficking in persons.

Despite this, as capacity and bandwidth to manage large amounts of quantitative data expand, so will the capabilities to draw in text-form data. There are already examples of narrative databases being developed for research use, although it is not yet possible to connect this form of qualitative data with quantitative indicators (unless both types of data were collected together). As systems of storing and reporting information improve, however, the quantitative indicators classified in the ICS-TIP can at some point be connected with this more complex information.

Hybrid approaches

While administrative data are clearly valuable on their own, they can be even more impactful when used in combination with other types of data, such as the ones described above. In other words, different data sources can be combined to balance each other's limitations and strengths. For example, the ICS-TIP can be used in some circumstances to combine sources of administrative data from different agencies and look for relationships between trafficking in persons and health, education or a number of other factors. From a broader perspective, administrative data could also be combined with certain types of survey data to provide a population estimate, or to ground-truth the geospatial data, correct/enhance the machine, and so on. A concrete example of this is the above-mentioned *Global Estimates of Modern Slavery* (see also Box 11 for more details).

However, linking data sets, and therefore more information, might also mean connecting more potentially identifying variables to cases. For this reason, all data must be presented in strict adherence with the same careful data protection standards discussed throughout the manual.

[72] Narrative, open-form text data can also be considered administrative data when they are collected by front-line agencies/government service providers, although most countries currently have limited capacity to collect and manage this amount of data.

Box 11. Global Estimates of Modern Slavery

In 2022, ILO, IOM and Walk Free Foundation published the second *Global Estimates of Modern Slavery*,[a] which provides a concrete example of how different data sources can be combined. The report found that on any given day in 2021, 50 million people were victims of modern slavery, including 28 million people in forced labour and 22 million people in forced marriage. In terms of prevalence, the report found that there were 6.4 victims of modern slavery for every thousand people in the world in 2021. How were these numbers obtained, given the limitations of administrative data and survey data?

They were obtained using a combined methodology, drawing on two different data sources. The first, and the central element, was specially designed, national probabilistic surveys involving interviews with nearly 72,000 and 110,000 respondents across 68 and 75 countries (for forced labour and forced marriage, respectively). The second data source was administrative data from IOM's database of assisted victims of trafficking. This second source was used specifically to estimate forced sexual exploitation, forced labour of children and the duration of forced labour exploitation.

In this case, the analysis combined the representativity of surveys together with the sample size and rich information on hard-to-reach populations provided by administrative data.

[a] See footnote 67.

USEFUL CONSIDERATIONS AND BEST PRACTICES FOR THE PRESENTATION OF ADMINISTRATIVE DATA TO VARIOUS AUDIENCES

Chapter V focused on different data users and their needs in terms of data to be shared. The stakeholders considered in this chapter are consumers of data presented by government agencies. Here the focus lies not on transfers of data, but on transmitting insights and analyses by presenting data in a manner consistent with their strengths, but also with their limitations, as described in the above section.

Too often, data presentation is limited to headline statistics that depict a phenomenon or the state of the world as if it were static. Headline figures in the form of point estimates can sometimes attract great attention when first released and can be a powerful means of advocating an accessible, data-driven message to a general audience. However, the presentation of descriptive statistics in the form of a single rate or count may be the least impactful way to convey all that the data can offer.

Knowing your audience

While headline statistics are designed to cast a wide net, reaching the broader general public, different target groups have different data communication needs and therefore will benefit from the analysis and presentation of richer, more descriptive detail. Administrative data have much to offer in meeting stakeholders' information needs (see Table 14).

Table 14. Possible information needs and suggested data presentation format by target group

Target group	Information needs	Data presentation format
Government agencies/ policymakers	National prevalence, patterns/ trends	Quick overview, infographic, graphs, headline figures
Local police	Mapping trafficking routes, at-risk industries and areas to monitor, recruitment patterns	Detailed information on perpetrators/recruitment process, presented geographically (private or public) and providing information on patterns[a]
Prosecutors	Details of cases, on trafficking networks and specifically perpetrators	Very specific information on cases/process (private information)
Assistance/front-line responders	Information (e.g. nationality) on how to assist or find vulnerable groups or victims and how to best serve them	Granular information on trends specific to area, type of exploitation, sector
International community	Indicators disaggregated by sex and age (national reporting on SDG targets, Global Compact for Migration commitments)	High-level overview, display progress/trends over time (public)
Researchers	Dynamic trends, interactions, causal relationships	Detailed information in interactive form, if possible (public, but more detail than headline figures)

a Such patterns should be interpreted with caution, as experts suggest there is a tendency to infer future (rather than current or past) trends from this information. The main point is that the data can be used to obtain descriptive rather than predictive information.

Once the target audiences and their unique information needs have been pinpointed, it is possible to select which details and relationships should be emphasized/prioritized for analysis and how best to present the findings to meet those needs. Keep in mind, also, that many relevant stakeholders will benefit much more from a local, rather than a national or global, view of the situation. Much of the response to trafficking in persons plays out at the local level and requires more detailed information from the ground up. Tapping into the detailed, context-specific information obtained from the analysis of administrative data, especially from the visual presentation of findings, can be highly impactful.

Relatedly, any presentation should be wary of how the target group might interpret the data and/or results presented. Simple explanations or examples of how to understand numbers and figures are often appreciated and can help prevent misunderstandings (particularly if there is no opportunity for a Q&A session with the presenter, as would be the case for a website or a report).

Assessing the confidentiality risks to avoid undermining data subject anonymity

Clearly, optimal presentation of administrative data varies depending on which type of data is being used and what insights the target group needs to draw. Knowing the target group will of course help to determine which data it is most useful to present but will also be critical to deciding which findings to share and which to restrict.

As described in Chapter IV, permission to access particular types of data is given on the basis of risk. The level of risk is determined by how sensitive the data are, and the level of clearance for different user types is dictated by the data-holding agency (specified in data governance protocols). The same principles of security that apply to personal, or very sensitive, data assets also hold true for the artifacts created from them. Recommendations for safe and effective presentation of data will therefore depend on whether the data and the products of data analysis are intended for internal (private) or external (public) use and the level of risk incurred.

It is good practice always to consider how the results of analyses, visualizations and other information communicated in the effort to combat trafficking in persons could be used in unintended ways. For instance, the kind of data presentation that can help law enforcement agencies identify victims or capture perpetrators could also be used by traffickers to find rescued victims or evade capture.

One frequent data protection challenge is when a subcategory consists of very few cases that could potentially reveal the identity of the individuals involved. For instance, if there are only three cases of TIP victims of a certain nationality exploited in a particular sector, it may be possible to identify any one of the data subjects with this information alone. Chapter V presented other such examples.

There are many instances in which the presentation of data characteristics or analysis findings can result in this kind of breach of privacy, endangering the data subjects. Examples include when:

- a very low number of cases is assisted by a particular CSO;
- a specific trafficking route with a small number victims is mapped, or data are reported at a granular location level;
- a graph is produced of the number of victims of different nationalities identified by a national referral mechanism and includes nationalities with very few cases;
- an interactive dashboard is used, in which users can select multiple characteristics and drill down to a small subset of individuals or an individual.

K-anonymization can safeguard against this kind of data protection breach. It involves setting a threshold for the minimum number of cases in any subcategory to be made public (for a more detailed description of the technique, see Chapter V), and can be crucial to avoid any data presentation or visualization in which the number of cases is so low that it may reveal the identity of data subjects.[73] Synthetic data can provide similar privacy guarantees (see Chapter V, section "Creating synthetic data").

Sharing the results of data analysis entails responsibilities similar to those arising from sharing data assets. An additional concern is the (unintentional) misrepresentation of findings. Careful attention to the guidance presented in this chapter will help to shed light on effective and safe data presentation and communication methods.

[73] This will always be the case for victims of human trafficking, though it may not be true for perpetrator data that can be public.

Knowing your data

Statistical and visual presentations of administrative data can offer crucial insights into what is needed to combat trafficking in persons, but can just as easily be used in ways that misrepresent or mislead, albeit unintentionally (a not uncommon occurrence).

To avoid this, it is important to understand the strengths and limitations of administrative data (described earlier in the chapter) and how they can and cannot be presented without overstepping the limitations. Two concrete (and frequent) examples where the limitations of TIP administrative data may be overstepped are (i) attributing changes in the number of recorded cases to changes in prevalence and (ii) overstating the representativity of administrative data. The rest of the section goes into greater detail on these particular pitfalls, starting with the former. Concrete examples of best practices are also provided in Annex 8.

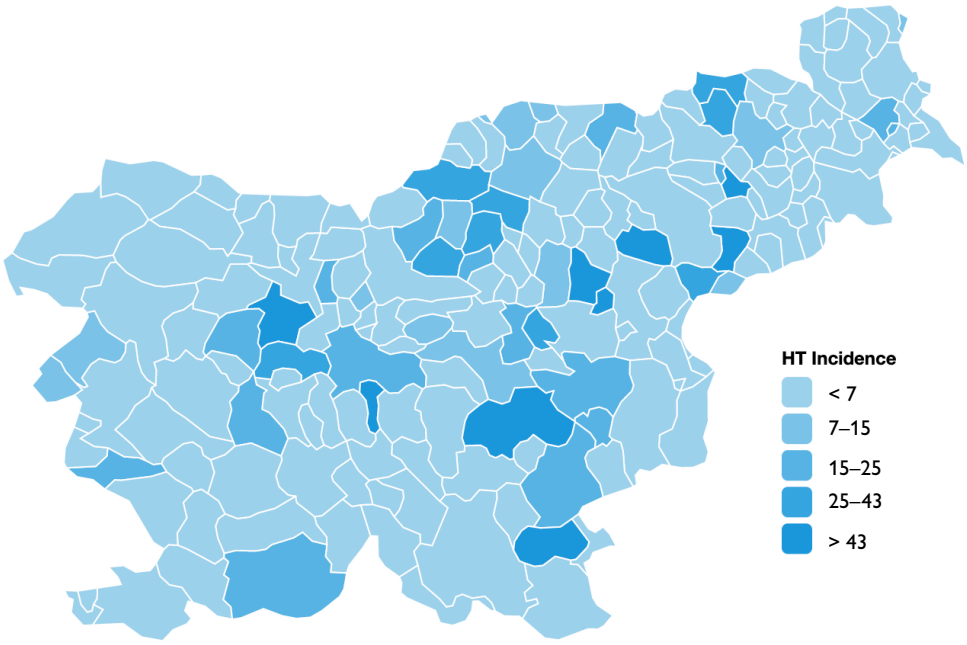
Attributing changes in the number of administratively recorded cases to changes in prevalence

Administrative data have much to offer when it comes to exploring patterns and trends in trafficking in persons, but they are not collected by randomly sampling a population with a relatively equal probability of being selected. The numbers of trafficking victims recorded often have more to do with how well data collection is going, rather than how many actual victims are out there. This means that, as data collection efforts improve, the number of identified victims will almost certainly increase, thus making it seem as if the TIP problem itself is growing, whereas the situation may instead be improving as governments become better at identifying and recovering victims.

An implication of this limitation is that it may be misleading to make comparisons of absolute numbers of recorded cases across localities or time using administrative data. As awareness of the issue grows and data-collection abilities improve the capture of this type of data (which is the aim of the ICS-TIP and this guidance manual), most countries will find the number of reported TIP cases rising – but that is not likely to be because the problem is getting worse, but instead because there is a greater focus on counter-trafficking efforts.

Presenting visual comparisons between localities, regions or countries can be a bigger problem depending on the purpose of the comparison. Case rates are typically mapped in different areas to demonstrate the need for assistance or law enforcement, or at least to draw attention to problem areas. However, if the areas with higher case rates are really those with better identification and services, the mapping is conveying the opposite message. This problem is illustrated by the choropleth map in Figure 17. Without any further information than what is on the map, viewers could believe that the areas that most in need of support are those in darkest blue. However, those in the lightest shade of blue may be those that are truly struggling to identify victims.

Figure 17. Choropleth map comparing incidence rates*



* Fictional map created by the project team.

Another major issue with comparison involves representing areas for which there are no data at all (because they do not exist or cannot be publicly accessed). In some instances, areas (countries, regions, municipalities) that are known to have serious issues with trafficking in persons may not collect or report the information. When they appear on a map next to other areas that are reporting cases, it looks as if they have no TIP problem at all. In Figure 17, there is no way of knowing whether the "fewer than 7" category includes areas which report no data at all, and why that might be the case. Fortunately, there is a simple, straightforward method of presenting missing data that avoids the error of presuming a lack of cases: clearly demarcating areas for which there are no data differently from areas with few or possibly no cases.

It is often more informative and less misleading to use relative rather than absolute values. For example, it may be possible and informative to investigate trends in the share of forced labour compared to sexual exploitation in a given country or region. Another example is to report the percentage of identified cases who are children or women on a choropleth map by region. Consider the concrete (but fictional) example in Table 15, summarizing the number of victims identified in different regions of a country and the number of children among them.

Table 15. Fictional example of a dataset

Region	Number of victims identified	Number of children	Number of adults	Per cent of children
Alpha	300	20	280	7
Beta	50	10	40	20
Gamma	100	10	90	10
Delta	250	15	235	6

Suppose that only the column with the raw number of children is then provided in a report. It might be an easy leap for a reader who is in a rush to infer that child trafficking is much more prevalent in Alpha than in Beta or Gamma, but that is not the full picture. Looking at percentages, 7 per cent of victims of trafficking in Alpha are children, against 20 per cent in Beta. This can suggest a very different situation.

However, even reporting data in percentages is not foolproof – what if there is a large CSO focused on child trafficking operating in Beta but not elsewhere, meaning that more children are identified in Beta? The next section looks at why the limitations of treating administrative data as representative of a whole TIP population must be assessed and clearly stated.

Overstepping the representativity of administrative data

Collecting data from front-line and other data-producing agencies will lead to information on individuals that would be much more difficult – or impossible – to reach by other means (traditional survey methods), and administrative data are considered in many contexts as the best source of information on the types of trafficking occurring at the national level (in the absence of an alternative).

However, it cannot be assumed that the entire, diverse population of TIP victims has been reached (see discussion on limitations to estimating prevalence with administrative data above). The section on the limitations of TIP administrative data pointed out that identified cases should be understood as a sample of the unidentified population of victims or perpetrators, yielding insight into trafficking trends and patterns. This sample may be biased if some types of trafficking cases are more likely to be identified (or referred) than others. Nevertheless, investigating this bias (or its absence) is challenging, given that unidentified individuals are, by definition, unknown. This potential issue is unlikely to affect all dimensions of the data equally. For example, while one may have confidence in the representivity of the age distribution of victims within the sectors of exploitation that are being identified, there may be concerns that some sectors are not being identified as often as they should.

One problematic implication of assuming that cases are nationally representative is that the harder-to-reach groups that may be most in need of interventions show up the least in the data. If it is assumed that low numbers of certain types of trafficking means that they are less common, when in reality it is because they may be more hidden and harder to identify, not enough effort will be made to assist those groups.

Relatedly, it is also important to be clear about where the data come from, specifically, their source, date and location of collection. Many important details are often left out in the interests of brevity and to reduce confusion or complexity, such as:

- Are they "known" cases of trafficking? Are they suspected or confirmed?
- Is the location listed (municipality, district, or other) the place where the event was reported, where it happened, or the location of the victim?
- Did the data come from a hotline, a police report, a criminal court case?

The strengths of administrative data, and the fact that in most cases there is no reasonable alternative, must therefore be balanced against the uncertainty regarding their representativity. In that respect, when presenting figures drawn from administrative data, it is important to be clear from the start that what is presented are considered "cases" or number of *detected* victims or perpetrators, and there may be limitations in terms of how representative the sample is of the whole TIP population, and in which dimensions.

ANNEXES

ANNEX 1. LEGAL BASES FOR PROCESSING PERSONAL DATA IN THE UNITED KINGDOM

Processing of personal data in the United Kingdom must be based on at least one of the GDPR Article 6 categories. An extract of Article 6 is copied below.[74]

Article 6

Lawfulness of processing

1. Processing shall be lawful only if and to the extent that at least one of the following applies:

   (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
   (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
   (c) processing is necessary for compliance with a legal obligation to which the controller is subject;
   (d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
   (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
   (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Point (f) of the first subparagraph shall not apply to processing carried out by public authorities in the performance of their tasks.

[74] The full text is available at www.legislation.gov.uk/eur/2016/679/article/6.

ANNEX 2. DIFFERENT DATA, DIFFERENT APPROACHES

An instructive example of the way legal frameworks handle different types of data is to be found in the comparison set out in Table A.2.1 below of how the United Kingdom GDPR views special category data and criminal offence data.[75] Further instructive examples are to be found in Australia's Privacy Act and Mexico's Data Protection Law.

Table A.2.1. Special category data versus criminal offence data in the United Kingdom

	Special category data	Criminal offence data
Definition	The United Kingdom GDPR defines special category data as: • personal data revealing **racial or ethnic origin**; • personal data revealing **political opinions**; • personal data revealing **religious or philosophical beliefs**; • personal data revealing **trade union membership**; • **genetic data**; • **biometric data** (where used for identification purposes); • data concerning **health**; • data concerning a person's **sex life**; and • data concerning a person's **sexual orientation**.	The United Kingdom GDPR Regulation gives extra protection to "personal data relating to criminal convictions and offences or related security measures". This is referred to as criminal offence data and covers a wide range of information about: • criminal activity; • allegations; • investigations; and • proceedings. It includes not just data which are obviously about a specific criminal conviction or trial, but also any other personal data relating to criminal convictions and offences, including: • unproven allegations; • information relating to the absence of convictions; and • personal data of victims and witnesses.
Rules	Data processing must be "generally lawful, fair and transparent" and comply with the United Kingdom GDPR. For processing to be lawful, an Article 6 basis must be identified.[a] Processing of special category data also requires meeting one of the specific conditions set out in Article 9. Five of the conditions are set out solely in Article 9. The other five require authorization or a basis in domestic law, which means that they must meet the additional conditions set out in the Data Protection Act of 2018.	Data processing must be "generally lawful, fair and transparent" and comply with the United Kingdom GDPR. For processing to be lawful, an Article 6 basis must be identified. Processing of criminal offence data also requires meeting one of the following condition: • it is undertaken under the control of official authority; • it is authorized by domestic law. This means that one of the conditions set out in Schedule 1 of the Data Protection Act of 2018 should be met. A comprehensive register of criminal convictions can only be kept if "under the control of official authority".

[75] See the website of the United Kingdom Information Commissioner's Office for more information on special category data and criminal offence data.

	Special category data	Criminal offence data
Conditions	Article 9 lists the conditions for processing special category data: (a) Explicit consent; (b) Employment, social security and social protection (if authorized by law); (c) Vital interests; (d) Not-for-profit bodies; (e) Made public by the data subject; (f) Legal claims or judicial acts; (g) Reasons of substantial public interest (with a basis in law); (h) Health or social care (with a basis in law); (i) Public health (with a basis in law); (j) Archiving, research and statistics (with a basis in law). The substantial public interest conditions number 23 and are set out in paragraphs 6 to 28 of Schedule 1 of the Data Protection Act of 2018. They include "[s]tatutory and government purposes", "[a]dministration of justice and parliamentary purposes", "[p]reventing or detecting unlawful acts", "[s]afeguarding of children and individuals at risk" and "[d]Disclosure to elected representatives".[b] Each condition has detailed provisions that must apply before it can be invoked.	The 28 conditions for the processing of criminal offence data are set out in paragraphs 1 to 37 of Schedule 1 of the Data Protection Act of 2018.[c] Conditions that can be relevant to the processing of TIP administrative data include: • Health or social care purposes; • Research; • Statutory and government purposes; • Administration of justice and parliamentary purposes; • Preventing or detecting unlawful acts; • Regulatory requirements relating to unlawful acts and dishonesty; • Safeguarding of children and individuals at risk; • Disclosure to elected representatives; • Judicial acts. Each condition has detailed provisions that must apply before it can be invoked.

[a] See Chapter III and Annex 1 for more information on Article 6, the full text of which is available at www.legislation.gov.uk/eur/2016/679/article/6.

[b] See the website of the United Kingdom Information Commissioner's Office for more information on the "substantial public interest" conditions.

[c] For more information on Schedule 1, see the website of the United Kingdom Information Commissioner's Office at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/criminal-offence-data/#schedule1 and www.legislation.gov.uk/ukpga/2018/12/schedule/1/enacted.

ANNEX 3. TWO EXAMPLES OF DATA PIPELINES

Table A.3.1 and Table A.3.2 were derived from Annex 1 and Annex 2.

Table A.3.1. Data pipeline – example 1

Data	Phase	Purpose and legal basis for processing data	Details
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Collection by front-line CSO	Informed consent obtained and legal basis for processing data established	The individual has been informed of: - the purpose of the interview; - the use to which the data collected in the interview will be put. The individual consented/assented to the interview. If the individual is a minor, the consent of the parent(s)/guardian(s) has been obtained. The guardian or appropriate authority was present during the interview, unless otherwise requested by the child.
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Internal operational use by CSO	Informed consent	The individual's full and informed consent has been obtained to conduct the screening interview to identify them as a victim of trafficking. This should be based on information given regarding the role of the organization, the voluntary nature of the interview and the use of the information provided by the individual. Informed consent is also necessary for all services, such as medical examination and procedure, health assessments, and immediate and long-term reintegration assistance.
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Referral to agency coordinating the national referral mechanism	Informed consent	The individual's full and informed consent has been obtained to share her/ his individual case data for assistance purposes with specified partner organizations involved in providing specified services.
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Referral to service providers	Informed consent	The individual's full and informed consent has been obtained to share her/ his individual case data for assistance purposes with specified partner organizations involved in providing specified services.
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Referral to police for formal identification and to explore possible support for investigation/ prosecution	Informed consent	The individual's full and informed consent has been obtained to share her/ his individual case data with law enforcement for formal identification and to explore possible support for investigation/ prosecution.

Data	Phase	Purpose and legal basis for processing data	Details
Derived from individual victim of trafficking, pseudonymized, no direct identifiers, but still sensitive as contains data related to a criminal offence and special categories of personal information (indirect identifiers)	Disaggregate, pseudonymized data sent to Ministry of Interior for reporting	Informed consent Control of official authority Substantial public interest	Use of anonymized data for: - archiving purposes in the public interest. - scientific research purposes. - statistical purposes. Police have authority mandated by law. Research: - archiving purposes in the public interest - scientific research purposes - statistical purposes Statutory and government purposes. Preventing or detecting unlawful acts.
Derived from individual victim of trafficking, pseudonymized, no direct identifiers, but still sensitive as contains data related to a criminal offence and special categories of personal information (indirect identifiers)	Aggregate data sent to national rapporteur	Control of official authority Substantial public interest	As above but Ministry of Interior has authority mandated by law.
Derived from individual victim of trafficking, pseudonymized, no direct identifiers, but still sensitive as contains data related to a criminal offence and special categories of personal information (indirect identifiers)	Aggregate data published on government open data portal or in report	Control of official authority Substantial public interest	As above but national rapporteur has authority mandated by law.
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	All front-line agencies notified by originating agency to dispose of data	Informed consent withdrawn	
Derived from individual victim of trafficking, pseudonymized, no direct identifiers, but still sensitive as contains data related to a criminal offence and special categories of personal information (indirect identifiers)	Ministry of Interior and National rapporteur maintain aggregate, pseudonymized data for reporting purposes	Informed consent withdrawn Control of official authority Substantial public interest	Use of anonymized data for: - archiving purposes in the public interest - scientific research purposes - statistical purposes

Table A.3.2. Data pipeline – example 2

Data	Phase	Purpose and legal basis for processing data	Details
	Police arrest group for criminal activity		
	Police identify victims of trafficking in group		
Provided by and pertaining to an individual victim of trafficking Includes personal data (direct and indirect identifiers), registration/ biodata, and data related to their needs and the criminal offence	Victims of trafficking agree to support police investigation and prosecution	Informed consent	The individual's full and informed consent has been obtained to make a statement to support police investigation and prosecution.
Includes indirect identifiers but not direct identifiers, data on the criminal offence, data on suspected victims and perpetrators involved	Prosecution reports suspected case to national statistics office	Control of official authority Substantial public interest	Prosecution has authority mandated by law. Research. Statutory and government purposes. Preventing or detecting unlawful acts.
	Court case concludes successful prosecution		
Includes indirect identifiers but not direct identifiers, data on the criminal offence, data on suspected victims and perpetrators involved	Prosecution and court report confirmed case to national statistics office	Control of official authority Substantial public interest	As above.
Includes indirect identifiers but not direct identifiers, data on the criminal offence, data on suspected victims and perpetrators involved	Aggregate data published on government open data portal or in report	Control of official authority Substantial public interest	As above.

ANNEX 4. TIPS ON ENCRYPTION

The standard type of secure encryption used to protect government data is the Advanced Encryption Standard (AES) set out in International Organization for Standardization/ International Electrotechnical Commission 18033-3, which specifies block ciphers for the purpose of data confidentiality. Encryption software with AES is publicly available and relatively easy to find and install.

One factor to note when setting up encryption to protect highly confidential records is that the key size used must be sufficiently long that the encryption code cannot be cracked by computers now or in the relatively near future. Cryptographic keys that use at least a 128-bit key length (the minimum AES standard) will be long enough to protect confidential data, although the most sensitive data typically requires 192 to 256 bits.

There are multiple ways to create encryption keys, including using Excel or free online software to generate a code.

ANNEX 5. THE THREE TYPES OF METADATA

Descriptive metadata

Descriptive metadata describe the basic assets of the indicators and dataset and should be as standardized as the data themselves. When government departments, such as law enforcement agencies, record a trafficking event, the corresponding metadata need to be created by the agency. In many cases, data on trafficking in persons are sourced from external data-producing agencies (for example CSOs, shelters, or other local government agencies). That means standard (simple) descriptive metadata must be required from those submitting data by the managing government department. Recording metadata for incoming data sourced from providers with the most limited capacity is also an option, although not recommended. A better solution would be to develop a streamlined form (digital or hard copy) that data-producing agencies can use to capture at least the most basic information about the data.

Providing detailed information on all aspects of the data once they have been collected is an essential step of data storage and management. Some datasets require highly technical metadata that provide an in-depth description of the sampling methods and the design and use of survey instruments. This level of description must be present in metadata on nationally representative survey data, for instance. For the purposes of guidance on metadata for the ICS-TIP, which is not survey data, but case data, this level of detailed information is not required. While not quite as complex, metadata on case data still needs to describe how the information was collected (see Box A.5.1). Many other basic characteristics of metadata are common to all data types as well.

Box A.5.1. Checklist of the minimum descriptive information to include*

Title
File name
Organization/department originally collecting the data
Date the data were submitted to repository
List of indicators
Number of cases
Contact of responsible party

* Metadata standard in social science: https://ddialliance.org/Specification/DDI-Codebook/2.5/.

Structural metadata

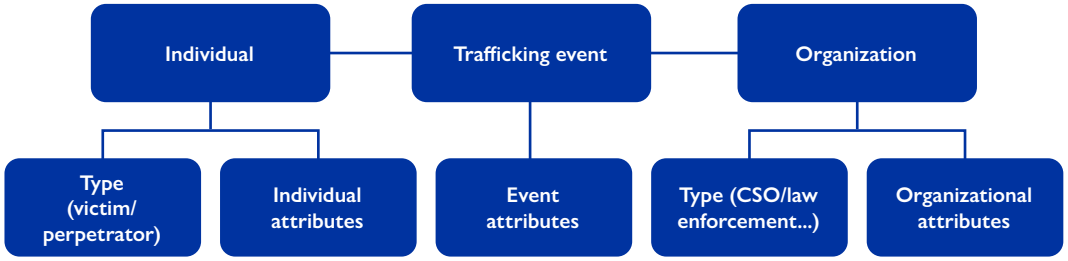
Data must be resolvable for both machines and humans. That means that data can be sourced and linked or connected to other datasets easily whether the process is conducted using code or by someone downloading the data and working with them following detailed instruction.

Structural metadata detail the format in which data are organized and addresses how they may fit with other types of data sources. Much of the information on how to organize and code indicators on trafficking in persons recommended by the ICS-TIP is set out in the ICS-TIP itself as well, in the form of a model identifying which indicators to collect data on, at what level of complexity, and with a system of numerical codes. Additionally, the ICS-TIP

provides information on how to match data on crime, labour, migration and health that are already classified indicators with the coding schemes of other international standards.

There are two types of documentation on organizing and storing data content in line with FAIR data principles and the most efficient processing of data by machines or individuals. First, the physical data model describes dataset contents and attributes, listing all possible values and how these indicators are measured. Second, the logical data model provides a mapping of how the indicators are stored and how each relates to the other (see Figure A.5.1).

Figure A.5.1. Example of a logical data model



Administrative metadata

Administrative metadata comprise the rules that govern storage and determine levels of access and protection for certain types of data. The answer to each of the questions in the above section on "Rules" will be set out here.

The guidelines for developing this core part of the data governance framework will generally follow the overarching principles listed in the introduction to Chapter IV. That is, the data governance plan developed by the government department overseeing data on trafficking in persons should be based on local and national legal norms, nuanced enough to manage diverse sources of data, apply the FAIR principles for maximum organization, and implemented by trained departmental officers.

Data management procedures such as how to classify different types of data according to level of security, where data should be stored and for how long will need to be formalized in this documentation.

ANNEX 6. THE MICROSOFT/IOM SYNTHETIC DATA ALGORITHM

The Microsoft/IOM algorithm works by creating a synthetic data set composed entirely of attribute combinations that are common in the sensitive data set (appearing at least *k* times). This means that (i) no rare combinations present in the sensitive data set are ever released in the synthetic data set, and (ii) any rare attribute combination present in the synthetic data set must, by construction, describe a larger group of individuals in the sensitive data set. In other words, no use of the synthetic data set permits discovery of groups smaller than *k* in the sensitive data set. For the CTDC, the value of *k* is 10.

The algorithm itself follows a multistage process to capture as much of the structure and utility of the sensitive data set as possible, while enforcing this privacy guarantee.

1. **Generate "seeded" synthetic records from the privacy-preserving core of each sensitive record.** For each sensitive record, create a synthetic record by randomly sampling attributes from the sensitive record (proportional to their joint frequency with attributes in the synthetic record that have already been sampled). Stop when it is no longer possible to sample another attribute without creating a rare combination, and track the attributes that could not be released.
2. **Generate "unseeded" synthetic records from the attributes suppressed during step 1.** Create synthetic records by randomly sampling attributes from the pool of unreleased attributes (proportional to their joint frequency with attributes in the synthetic record that have already been sampled). Continue until all attributes have been accounted for, even if it means releasing synthetic records with single attributes.
3. **Suppress attributes to avoid disclosing precise attribute counts.** Steps 1 and 2 perfectly preserve the counts of individual attributes appearing at least k times in the sensitive data set. Randomly suppress instances of each attribute until its count reaches the next lowest multiple of k.
4. **Sort synthetic records by length to intermix seeded and unseeded variants.** As a final step, the synthetic records of the output data set are sorted by decreasing count of non-null attributes. This removes any relationship to the order of sensitive records and the distinction between seeded and unseeded variants.

Figure A.6.2 provides a concrete example of how the algorithm works (steps 1 and 2 described above).

The data are processed in a way that preserves not only the structure of the data set but, more importantly, the statistical relationships between attributes. This "seeded synthesis" approach thus keeps attributes connected, safely, without exposing any unique combination of attributes that could be used for re-identification. It also ensures that rare attributes are not "dropped" out of the newly generated data.[76] From each sensitive record, the longest subset of attributes that describe a "common" attribute combination is extracted and can thus be safely released. Any remaining attributes from the sensitive record are tracked and later combined into new records.[77]

[76] To be clear, the Microsoft/IOM approach never drops records – it drops attributes from records and then combines them to the greatest extent possible so that they may still be released, even if this means releasing records with just a single attribute.

[77] The CTDC publishes synthetic data generated in a seeded way. However, the synthetic data showcase also allows data to be generated in an unseeded way by randomly sampling joint attribute distributions. Both methods are privacy preserving but the resulting data structures are different. See https://github.com/microsoft/synthetic-data-showcase for more information.

Clearly there are numerous advantages to using this method of creating synthetic data, though, as with any approach, there are some limitations. As explained earlier, synthetic data can only preserve both privacy and utility when the dimensionality of the data is low. In other words, the procedure will be more useful when the data set is less likely to hold a large number of rare groupings of attributes. Indeed, if the data have many unique or rare combinations of attributes, the sensitive records will need to be "broken up" more to hide these potentially identifying combinations. The synthetic data set then ends up with many more records than the sensitive data set (since these records are necessary to preserve univariate attribute counts). This process depends on the *k* parameter chosen (as the approach satisfies the k-anonymity property): a lower *k* may yield fewer "broken down" records but will pose a greater privacy risk. The process also depends on how many attributes are included in the synthesis. Indeed, the more traits are combined in a single record, the greater the potential that the record in question will contain a uniquely identifying attribute combination (and will in turn have to be "broken up"). Some workarounds may exist, such as splitting the data set to create subsets of synthesized data on a smaller set of attributes. Figure A.6.1 provides a simple example of a case where records have to be broken up, while Figure A.6.2 gives a more detailed and concrete example of how a record is generated by the algorithm.

Figure A.6.1. Simple example of the creation of a new synthetic record for *k*=2

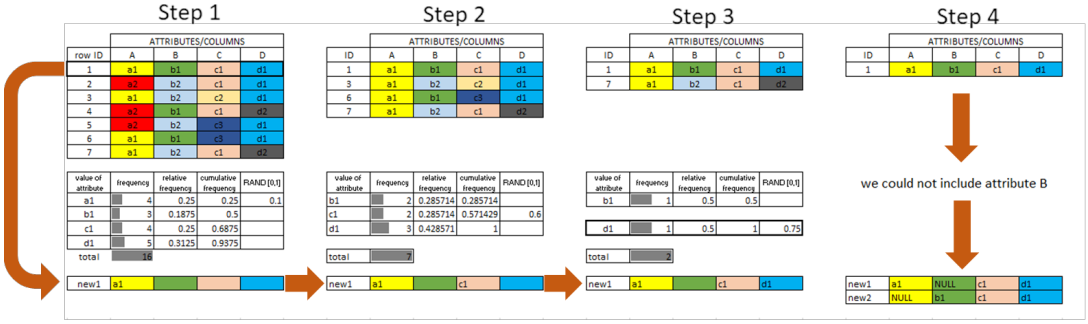
Real data

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

Synthetic data

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

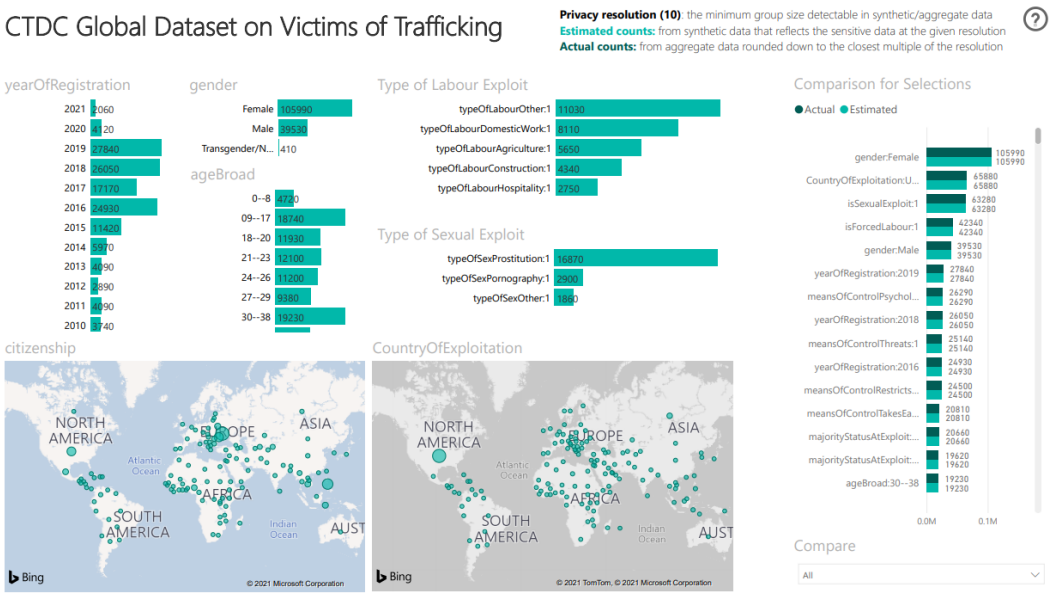
Figure A.6.2. A concrete example of how the Microsoft/IOM algorithm works



Source: created by Eduardo Zambrano, Displacement Tracking Matrix, IOM.

In addition to generating a synthetic data set, the algorithm published by Microsoft automatically generates interactive Power BI data dashboards, which allow users not only to interact easily with the synthetic data set, but also to compare it (to the extent possible) to the original data, thanks to pre-computed aggregate counts. Figure A.6.3 provides an overview of such a dashboard.

Figure A.6.3. Screenshot of a Power BI dashboard generated by the Microsoft/IOM algorithm



ANNEX 7. ESTIMATING PREVALENCE USING MULTIPLE SYSTEMS ESTIMATION

While administrative case data can be used to determine the characteristics of and obtain insights into the crime of trafficking in persons and the persons, organizations and locations involved, they are less useful for estimating the national prevalence of trafficking in persons, because the data are obtained without probabilistic sampling.

However, certain statistical techniques can be used to estimate national prevalence by combining sources of administrative data if certain methodological criteria are met. The chief such technique is multiple systems estimation, or MSE. MSE is a statistical approach used to estimate the size of hidden populations. Adapted from a long-standing probabilistic method used to estimate hard-to-reach populations known as "capture-recapture", MSE requires lists of subjects gathered from multiple sources (three or more, ideally) that can be compared to identify which subjects where captured each time and who seemed to be missed.

Accounting for complex social characteristics and behaviour, especially when a crime is involved, complicates the process of identifying victims. MSE has been adapted from capture-recapture methods to deal with some of this complexity by varying the statistical modelling used to produce estimates. UNODC, for example, has supported estimates of national prevalence in several countries, pioneering efforts to use administrative case data for MSE in the Kingdom of the Netherlands.[78]

Several requirements must be met when using MSE to obtain national prevalence rates. The first, key element to identifying cases arising on more than one list lies is each data source's unique identifiers. In other words, raw case data are needed that can be matched to find which individuals were recorded on more than one list. Secondly, multiple lists (at least three) are needed to find the overlaps. Thirdly, the lists must reasonably cover all areas (geographically and by type of exploitation) and not leave out areas with less coverage by CSOs or types of trafficking that are not as commonly detectable.

In order to make inferences about national prevalence based on the estimates, a number of assumptions of the MSE model must be met, namely:

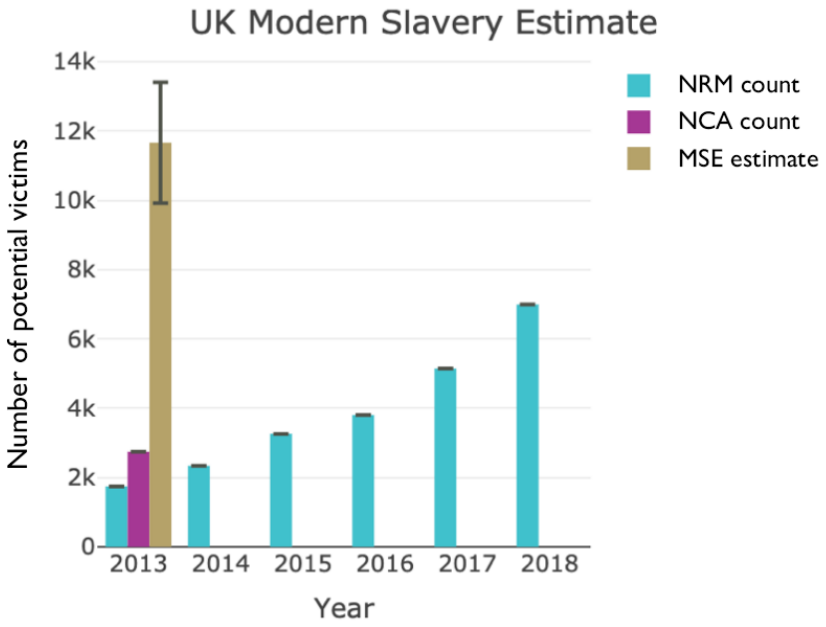
1. Closed system: the population does not change (increase or decrease) during measurement period;
2. Independence: being recorded on one list does not affect the probability of being on another list;
3. Homogeneity: every individual has a non-zero probability of being selected (or at least not systematically excluded).

The model has several notable limitations as well. The first, which has already been touched on, is that MSE is not helpful in all contexts, especially not in those where few data are collected and there is no guarantee of comprehensive lists or lists with many overlapping cases. A second, related, point is that MSE cannot solve all issues in terms of finding deeply hidden subsegments of the population that may never appear on any list. Thirdly, MSE offers an estimate range rather than a point estimate (as is true of estimates from survey data as well), which is only useful if the range is not too large. Finally, a major drawback lies in the need for data with unique identifiers that can create data protection issues, allow

[78] See www.unodc.org/documents/data-and-analysis/tip/TiPMSE.pdf.

only certain researchers access and do not allow the data to be shared for the purposes of replication or verification.

Figure A.7.1. Comparison of the MSE estimate in the United Kingdom with national referral mechanism count



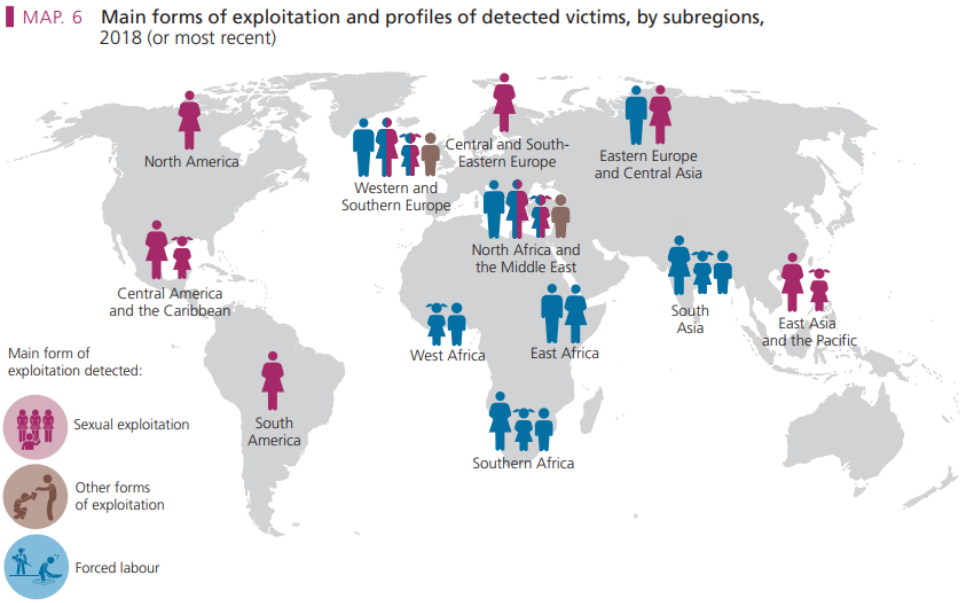
Source: Chart generated by the core team using data from the 2014 MSE carried out in the United Kingdom (available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf).

Note: NCA stands for National Crime Agency, which, before April 2019, was responsible for collecting data on the national referral mechanism (see www.nationalcrimeagency.gov.uk/what-we-do/crime-threats/modern-slavery-and-human-trafficking).

As is illustrated in Figure A.7.1, the data obtained from an MSE should be visualized in line with a number of good practices. First, the difference between the number of cases detected and the estimated total population of cases should be displayed, to highlight the point that the number of detected cases is likely much lower than the actual number of cases. The relationship between detected and estimated cases over time should also be displayed. While the number of detected cases will likely grow as data collection improves, the number of undetected cases should shrink, as the gap between detected and undetected cases closes. Lastly, the estimate range should be displayed alongside the point estimate, so as to indicate that the total number of cases is an estimate and how accurate the estimate is likely to be.

ANNEX 8. CONCRETE EXAMPLES OF DATA PRESENTATION

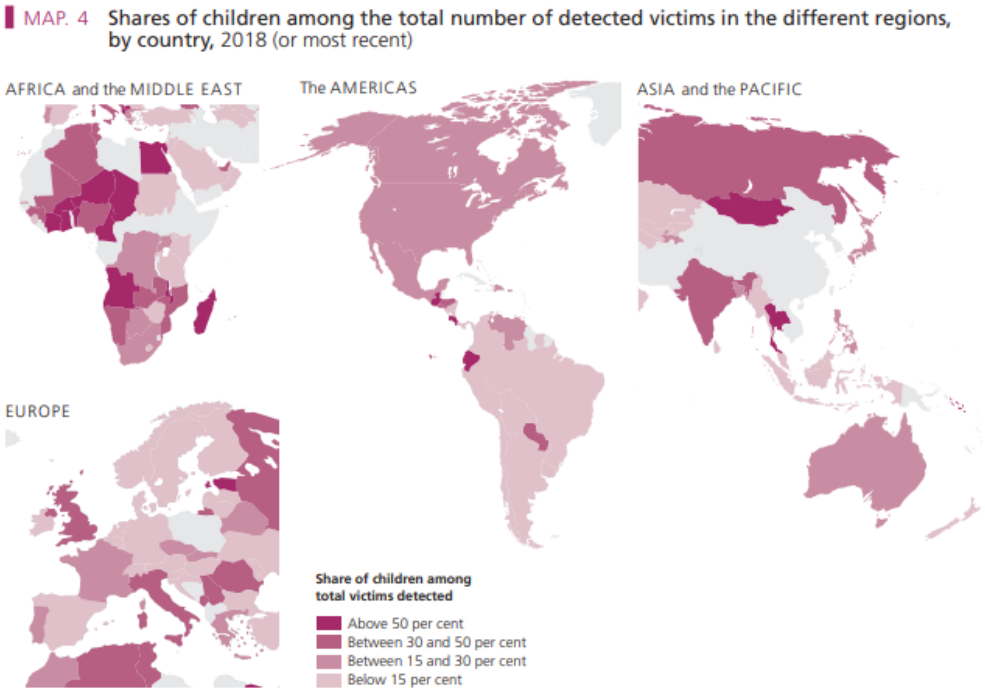
Figure A.8.1. Map on profiles of exploitation and gender



Source: UNODC, *Global Report on Trafficking in Persons 2020* (United Nations, New York, 2020), p. 37.
Note: This map is for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by the International Organization for Migration.

The map shows the most frequent profiles of exploitation and gender per region. It avoids providing absolute figures or indicating the region with the highest number of each type of exploitation. It also provides the source of the data, at the bottom left, and a date. The same goes for the map below.

Figure A.8.2. Regional maps on share of children detected

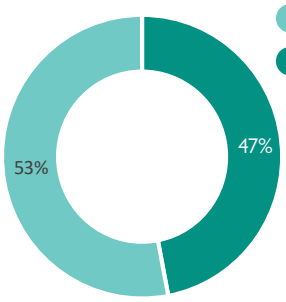


Source: UNODC, *Global Report on Trafficking in Persons 2020* (United Nations, New York, 2020), p. 32.
Note: These maps are for illustration purposes only. The boundaries and names shown and the designations used on these maps do not imply official endorsement or acceptance by the International Organization for Migration.

Notice also how the map clearly displays countries for which no data are available.

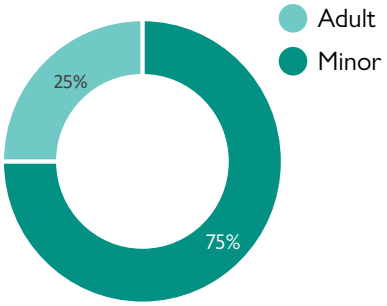
Figure A.8.3. Graphs comparing the share of children in sexual exploitation

Majority status of men and boys trafficked into sexual exploitation



In CTDC data, just over half (52%) of all male victims traficked into sexual exploitation are children.

Majority status of all victims trafficked into sexual exploitation

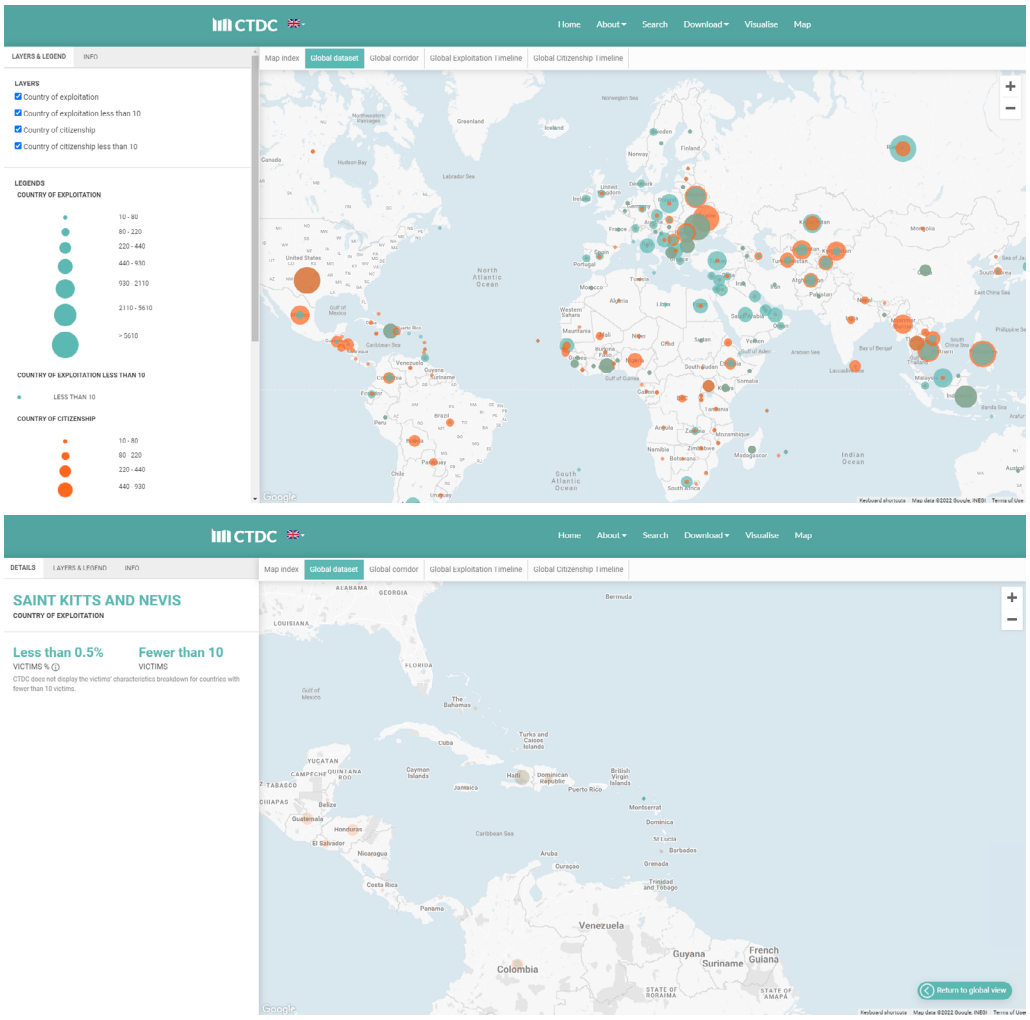


In comparison, the same number for all victims trafficked into sexual exploitation is 25%.

Source: CTDC, Men and boys trafficked into sexual exploitation (accessed January 2022).

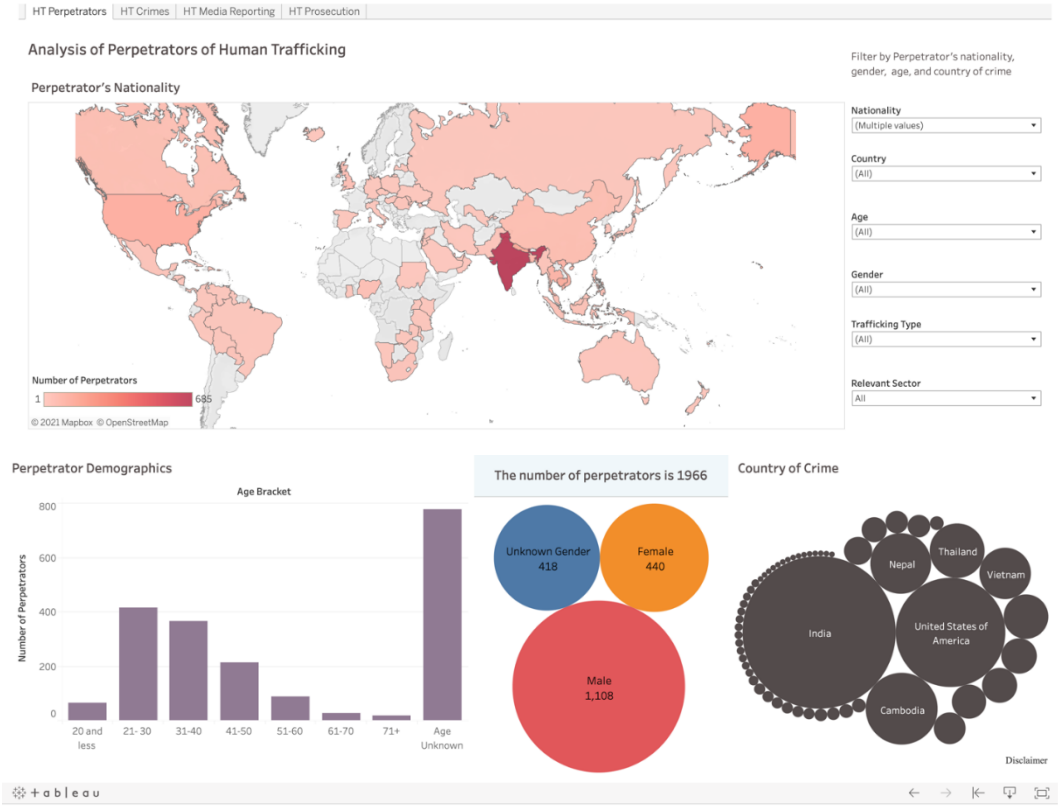
The graphs above provide an age comparison for men and boys trafficked for sexual exploitation. This is an example where relative values (here, percentages) give a much clearer picture than absolute numbers. Notice the sentence underneath each graph, explaining to readers how to interpret the figures provided.

Figure A.8.4. Screenshots of one of the CTDC's map

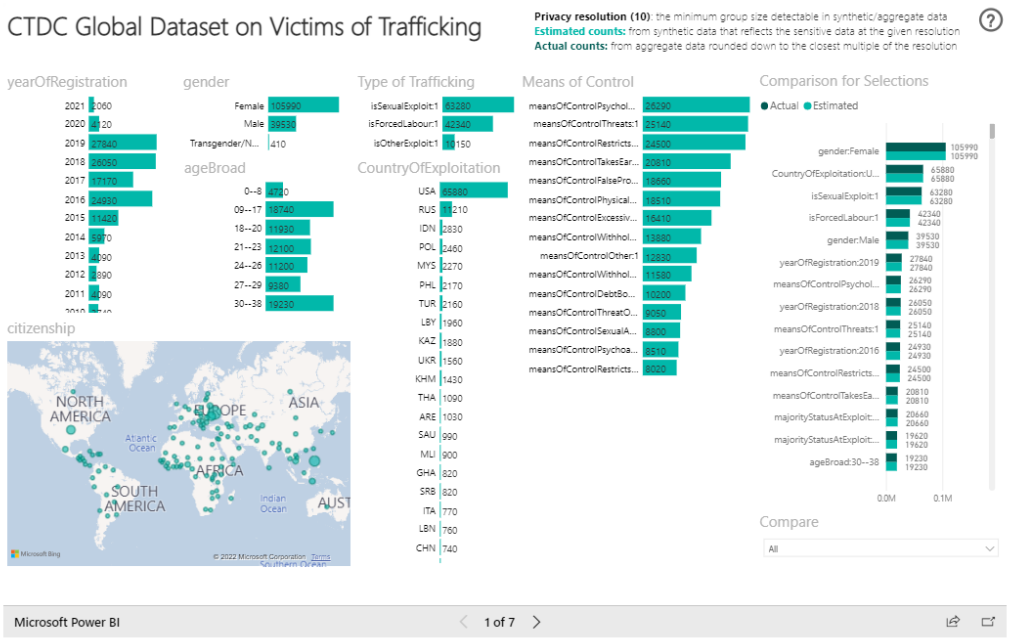


Source: CTDC (accessed January 2022).
Note: These maps are for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by IOM.

Figure A.8.5. Examples of interactive dashboards



Source: Liberty Shared (accessed January 2022).



Source: CTDC (accessed January 2022).
Note: These maps are for illustration purposes only. The boundaries and names shown and the designations used on this map do not imply official endorsement or acceptance by IOM.

The main pitfall to avoid in this case is ensuring users cannot drill down to one or a few individuals. This can be avoided by preparing pre-computed averages to power the dashboard (as in the case of the CTDC dashboard). It may also not be necessary, depending on the data. In the case of the Liberty Shared dashboard, the data are already publicly available.[79]

[79] See https://libertyshared.org/idc-center for more explanations.

International Organization for Migration (IOM)
17 route des Morillons, P.O. Box 17, 1211 Geneva 19, Switzerland
Tel: +41 22 717 9111 • Fax: +41 22 798 6150 • Email: hq@iom.int • Website: www.iom.int