OCHA centre for humdata

THE CENTRE FOR HUMANITARIAN DATA

GUIDANCE NOTE SERIES
DATA RESPONSIBILITY IN HUMANITARIAN ACTION

NOTE #1: STATISTICAL DISCLOSURE CONTROL

KEY TAKEAWAYS:

- Statistical disclosure control (SDC) is a technique used to assess and lower the risk of a person or organisation being re-identified from the analysis of microdata.
- In the humanitarian sector, microdata is data on the characteristics of a population that is gathered through exercises such as household surveys, needs assessment or monitoring activities.
- The purpose of applying disclosure control to humanitarian microdata is to be able to share the data more widely in a responsible manner without harming affected people.
- An SDC process can lower the risk of re-identification to an acceptable level but the risk threshold may vary depending on the context where the humanitarian response is happening.
- To start using SDC, organisations should invest in (1) finding the right tool, (2) setting up a workflow, and (3) improving practice over time through continuous learning.

WHAT IS HUMANITARIAN MICRODATA?

Data on the characteristics of units of a population (e.g. individuals, households or establishments) collected by a census, survey or experiment is referred to in statistics as 'microdata'.[1] In humanitarian response, this type of data is gathered through exercises such as a Multi-Sector Needs Assessment (MSNA), household surveys, and other needs assessment or monitoring activities. Such data make up an increasingly significant volume of data in the humanitarian sector, and are evermore critical to determining the needs and perspectives of people affected by crises.[2] As such, it is essential that humanitarian organisations understand how to assess and manage the sensitivity of this data in order to ensure its full use and impact in different response contexts.

In its raw form, microdata can contain both personal data and non-personal data on a range of topics, including sensitive subjects such as exposure to gender-based violence, infectious diseases, and other issues that may be recorded in free text fields. Most humanitarian organisations acknowledge the sensitivity of personal data such as names, biometric data, or ID numbers and anonymise data sets accordingly as a matter of standard practice. However, it is often still possible to re-identify individual respondents or organisations by combining answers to different questions, even after such 'anonymisation' is applied.

[1] Survey Design and Statistical Methodology Metadata, Software and Standards Management Branch, Systems Support Division, United States Bureau of the Census, Washington D.C., August 1998, Section 3.4.4, page 39.

[2] At the time of writing, a search for the word 'survey' on the Humanitarian Data Exchange returned 1198 results out of the 9805 datasets on the platform; a search for the word 'assessment' returned 1399 results.

RE-IDENTIFICATION AND DISCLOSURE RISK

A string of data points can allow for re-identification, either in isolation or when combined with basic contextual understanding. Advanced data analysis techniques can also extract more sensitive insights than may be visible through basic analysis, increasing the potential sensitivity of microdata in the humanitarian sector.

There are three commonly recognised forms[3] of disclosure risk, each of which could manifest in humanitarian microdata:

- **Identity disclosure:** when a known individual can be associated with a released data record

- **Attribute disclosure:** when some new characteristic of an individual can be determined based on the information available in the released data

- **Inferential disclosure:** when some characteristic of an individual can be determined more accurately with the released data than would otherwise have been possible

STATISTICAL DISCLOSURE CONTROL

One method for assessing and reducing the risk of re-identification in microdata is Statistical Disclosure Control (SDC). SDC is a technique used in statistics to assess and lower the risk of a person or organisation being re-identified from the results of an analysis of survey or administrative data, or in the release of microdata. This technique has primarily been used by National Statistical Offices (NSOs) and other statistical organisations in relation to census data.

By applying SDC, the overall informational value of a dataset will always be impacted, and striking an appropriate balance between re-identification risk and loss of information value is critical to any application of SDC. In determining the appropriate risk-utility balance, it is essential to account for the various possible uses of a dataset and the context in which the data was collected.

Typically, a process around SDC will consist of three steps:

1. **Risk assessment**
   In the first step, assess the probability that successful disclosure could occur for individual respondents within a given dataset. Whether a risk percentage is acceptable for a dataset will depend on the context to which the data relates. For example, in a conflict environment, the permissible risk percentage will typically be lower than in a natural disaster response.

2. **Application of Statistical Disclosure Control methods**
   In the second step, put the dataset through the actual SDC process, which lowers the re-identification risk by applying one or more methods for anonymisation. These methods fall into one of two categories: perturbative methods, which do not suppress values in the dataset but perturb (i.e., alter) values to limit disclosure risk by creating uncertainty around the true values, or non-perturbative methods, reduce the detail in the data by generalisation or suppression of certain values (i.e., masking) without distorting the data structure.

3. **Reassessing risk and quantifying information loss**
   The final step in the process is to measure the information loss resulting from the different SDC methods applied to the dataset. This compares the information value of the original dataset with the final information value, and also entails reassessing the risk of re-identification to ensure that it has been sufficiently reduced.

[3] For more information on disclosure risk and related technical considerations for the assessment and management of such risk through SDC, see **the Statistical Disclosure Control for Microdata: A Practice Guide, available here: https://sdcpractice.readthedocs.io/en/latest/index.html.**

The purpose of applying SDC to microdata in the humanitarian sector is to be able to share survey and needs assessment data more widely in a responsible manner. Using SDC as a means for assessing and reducing the sensitivity of data, humanitarian organisations can responsibly disseminate survey and needs assessment data to inform the overall response effort.

Applying SDC to data shared on HDX

Since the beginning of 2018, the Centre for Humanitarian Data has conducted a risk assessment of 59 datasets uploaded to the Humanitarian Data Exchange (HDX) platform. The risk of disclosure of respondents' identities in 38 of those datasets was considered too high for publication on HDX. The contributors of 14 of these datasets agreed to the application of SDC to their data to lower the risk level. The HDX team applied SDC to these 14 datasets, for which the risk level was lowered to an acceptable level (i.e. 5% or lower[4]). For 5 of these 14 datasets, this meant that they could be made public after anonymisation. The remaining 9 datasets were either removed or shared privately on HDX, as were the 24 high-risk datasets for which the contributor did not agree to conduct SDC. For those 24 datasets, many organisations took measures of their own to lower the risk of re-identification, sometimes including the removal of non-essential sensitive variables altogether.

APPLICATIONS OF SDC IN HUMANITARIAN DATA MANAGEMENT

In early 2019, the Centre for Humanitarian Data conducted interviews with seven humanitarian organisations that conduct regular surveys and needs assessments to understand existing practices for microdata management. While some organisations such as UNHCR (see case study below) have relatively advanced approaches and considerable expertise in-house for conducting SDC on different forms of microdata, most of the organisations interviewed require support to do this work.

Responsible Curation and Management of Microdata about Refugees

Experience from UNHCR

UNHCR routinely collects data on refugees and other populations under its mandate. This data is used to assess needs and vulnerabilities, inform programming and better target assistance. Although this data has not traditionally been retained in formats and locations that would make it easily retrievable for future use, UNHCR is now in the process of creating an internal and an external microdata library. By creating these online repositories that will allow public access to microdata for internal and external users, UNHCR aims to enable more extensive use of the data by a variety of stakeholders and prevent duplication in data collection efforts moving forward.

While public dissemination of microdata has many potential benefits, it also comes with potential risks. Dissemination without appropriate disclosure control measures can enable intruders to identify the entities (individuals or households) whose data is being shared, even if direct identifiers like names and addresses have been removed. In accordance with **UNHCR's data protection policy**, the identity of persons of concern must be protected, and therefore datasets must be properly anonymized before they can be shared. UNHCR data is especially sensitive, as it concerns particularly vulnerable groups of people, whose protection is of the utmost importance.

To ensure protection and responsible dissemination of microdata, UNHCR utilizes the sdcMicro app in R to calculate the re-identification risks of such data before they are published. The process is managed by UNHCR's data curation team, which works together with the data owners to identify key variables, assess the sensitivity of the data, and set an acceptable risk level for a particular dataset. After anonymisation, the modified data is compared to the original and assessed for information

[4] The Centre recently adjusted its default threshold of acceptable reidentification risk from 5% to 3%. The exact threshold for a particular dataset is always contextual and is determined together with the organisation contributing the dataset

loss. If the data owner judges that certain modified variables are essential for consumers of the data, the disclosure control methods can be adjusted accordingly. For example, in the case of the Standardized Expanded Nutrition Surveys (SENS), the curation team decided not to apply aggregation in age brackets that would normally be applied because these brackets were key to characterizing malnutrition by age in years and months for children. The team maintained the age variable but excluded the date of birth and the survey date. This led to an acceptable risk scenario while keeping the data useful for nutritionists.

UNHCR continues to invest in this process by growing its curation team and increasing the technical expertise in anonymisation techniques within the organisation. Under the current plan, the UNHCR Microdata Library will be fully operational and populated with forced displacement microdata at the end of 2019.

By working with data contributors like REACH (see case study below) to develop a reliable and secure SDC service, the Centre for Humanitarian Data aims to support responsible sharing of this data and demonstrate the value of more robust techniques for risk assessment and data anonymisation. Exposure to these techniques is helping different humanitarian organisations to identify tools and methods that they can incorporate into their own data management processes, while also contributing to the broader body of knowledge within the sector on how to more responsibly manage and share microdata in humanitarian contexts.

Opportunities and challenges to incorporating SDC into an organisation's workflow

Experience from REACH

REACH began exploring the potential of SDC in June 2018, when the HDX team first applied the sdcMicro R package to a dataset that REACH uploaded to the platform. The types of data for which the HDX team have applied SDC for REACH include household surveys and key informant interviews (and associated metadata). REACH has not yet applied SDC directly but are looking into the requirements for doing so.

Based on experience to-date, REACH suggests that organisations interested in incorporating SDC into their workflow consider the following questions:

- Is this the right methodology for your existing microdata management processes?
- To what extent does application of SDC lower the validity and utility of the data?
- How does the application of SDC affect transparency?
- How can you ensure that personnel do not rely too much on the outcomes of an SDC disclosure risk assessment, and ensure that they keep thinking critically about potential risks of different data types?

REACH has determined that it would be operationally feasible to roll out the technical aspects of SDC relatively easily both at HQ and field level. At HQ this would mean running a script on all datasets produced or published by REACH. At the field level, this would mean getting country teams to use sdcMicro or a similar tool on all datasets produced in country.

Beyond the technical aspects of SDC, REACH sees the potential challenge or bottleneck in the manual component of the process whereby staff must decide whether a particular disclosure control technique is appropriate, which variables to remove or otherwise obfuscate, and how to interpret and communicate the results of the process. These decisions take time and require an understanding of the context to which the data relates.

In the near-term, REACH will continue collaborating with the HDX team to conduct SDC on survey and assessment data before publication on HDX. This experience will enable REACH to determine how best to incorporate SDC into its own workflows at the global and country level in the future.

AREAS FOR INVESTMENT: INCREASING THE USE OF SDC IN HUMANITARIAN MICRODATA MANAGEMENT

Investing in the capacity and infrastructure required to run SDC will allow organisations to determine the risk associated with sharing survey and needs assessment datasets. The Centre recommends the following three areas for investment for successful adoption of SDC by humanitarian organisations.

A. Tools

Various organisations have developed open source tools to conduct SDC and have made them available for free online. The Centre for Humanitarian Data and other humanitarian organisations consulted during the Centre's research currently use sdcMicro. The Centre chose sdcMicro for its scalability and because it is free and open source.

Other free and open source tools include μArgus and ARX. In selecting the appropriate tool for conducting SDC, organisations should consider the capacity of the tool to handle large metadata, the risk-utility trade off, and the ease with which staff would be able to navigate the tool's user interface.

B. Workflow

Besides selecting the right tool for SDC, setting up a clear workflow for application of the selected tool is key. SDC requires various roles to be involved throughout the process, including a technical specialist to apply the tool, someone with an understanding of the context to which the data relates to determine the acceptable risk level, and others. A well-organised workflow will help improve efficiency of the process and help prevent misinterpretation of or overreliance on the outcomes of SDC.

C. Continuous Learning

As organisations apply SDC, they will learn over time about the sensitivity of different key variables, the appropriate risk level to strive for, the acceptable level of information loss, and various other considerations that must be balanced in the process. Keeping a record of each application of SDC and documenting lessons learned is important, especially in the initial period when SDC is being introduced. Sharing these insights internally across teams and, as appropriate, with the broader humanitarian community is a great way to support more consistent and responsible management of microdata in the sector.

As part of its efforts to support more responsible management and sharing of sensitive humanitarian data, the Centre for Humanitarian Data is enhancing its current service model for conducting SDC. This work includes the introduction of an automated risk detection process for all data shared through HDX, which -- when done manually -- can take several hours for large spreadsheets. Through this process, a script will run on all data uploaded to the platform to identify microdata and other forms of potentially sensitive data. High-risk data will be sent into a dedicated workflow to be assessed and, if necessary, modified through SDC to reduce re-identification risk before the data is shared more widely.

To learn more about the Centre's work on SDC for humanitarian microdata, contact centrehumdata@un.org.

COLLABORATORS: UNHCR, REACH INITIATIVE

The **Centre for Humanitarian Data**, together with key partners, will publish a series of Guidance Notes on Data Responsibility in Humanitarian Action over the course of 2019 and 2020. The Guidance Note Series follows the publication of the **working draft OCHA Data Responsibility Guidelines** in March 2019. Through the series, the Centre aims to provide additional guidance on specific issues, processes and tools for data responsibility in practice. This series is made possible with the generous support of the Directorate-General for European Civil Protection and Humanitarian Aid Operations (DG ECHO).

This document covers humanitarian aid activities implemented with the financial assistance of the European Union. The views expressed herein should not be taken, in any way, to reflect the official opinion of the European Union, and the European Commission is not responsible for any use that may be made of the information it contains.

This project is co-funded
by the European Union