STATS BRIEF

Exploring new horizons: Nepal's pilot in implementing data integration approaches

May 2023 | Issue No. 33

UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific

Disclaimer: The designations employed and the presentation of the material in this Stats Brief do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. Where the designation "country or area" appears, it covers countries, territories, cities or areas. Bibliographical and other references have, wherever possible, been verified. The United Nations bears no responsibility for the availability or functioning of URLs. The opinions, figures and estimates set forth in this publication should not necessarily be considered as reflecting the views or carrying the endorsement of the United Nations. The mention of firm names and commercial products does not imply the endorsement of the United Nations.

This Stats brief is prepared by Jose Ramon Albert, Hossein Hassani, and Ram Hari Gaihre, the consultancy team of the project, under the supervision of Afsaneh Yazdani, Statistician, ESCAP Statistics Division, and Petra Nahmias, Chief of Population and Social Statistics Section, ESCAP Statistics Division. The authors prepared this Stats Brief based on the report titled "Strengthening the National Statistical Capacity to Implement Data Integration Approaches".[1] *The authors extend their gratitude to the National Statistics Office of Nepal, particularly Munni Kumari Chaudhary, Deputy Chief Statistician of the Planning and Human Resource Management Division, and Rishi Ram Sigdel, Director of the Training Section, for their generous support throughout the project. The authors are also grateful to Sovannaroth Tey for his invaluable input and contribution to the project.* For further information on this policy brief, please address your enquiries to:

Rachael Beaven
Director, Statistics Division
Economic and Social Commission for Asia and the Pacific (ESCAP)
Email: stat.unescap@un.org

Tracking number: May 2023 | Issue No. 33

Photo credit: iStock-849234128.jpg

1 https://www.unescap.org/kp/2023/strengthening-national-statistical-capacity-implement-data-integration-approaches-pilot.

Table of Contents

List of boxes

Summary

Data integration approaches are becoming increasingly popular among official statisticians and data producers; however, more institutional and technical capacity is required to fully realise their potential. To strengthen the capacity of statistical systems in Asia and the Pacific, ESCAP has released the Asia-Pacific Guidelines to Data Integration for Official Statistics[2] and launched the Data Integration Community of Practice, [3] a forum for sharing relevant knowledge and experience and learning. Furthermore, in 2022, ESCAP, in collaboration with the National Statistics Office (NSO) of Nepal, initiated a project to implement a tailored approach in Nepal to improve the capacity of the National Statistical System (NSS) to integrate data and produce better data for measuring Sustainable Development Goal (SDG) indicators. This Stats Brief provides an overview of the project and the technical advisory and capacity-building support provided to the NSO of Nepal and its data partners.

2 https://www.unescap.org/kp/2021/asia-pacific-guidelines-data-integration-official-statistics.

3 https://www.unescap.org/our-work/statistics/communities.

Introduction

National Statistical Offices (NSOs) around the world are extending their boundaries in the production of statistics in response to the changing context in which they operate. Global commitments, such as the 2030 Agenda for Sustainable Development, and national development plans have significantly raised the demand for quality, timely, and granular statistics. NSOs are responding to budget constraints and declining response rates by moving away from traditional data collection methods, including censuses and sample surveys, towards making better use of administrative data as well as new data sources. With technological advancements and the availability of new data sources, NSOs now have more options at their disposal.

In Asia and the Pacific, according to the recent SDG Progress Report, 128 out of 231 SDG indicators have sufficient data availability. The report indicates that *the lack of sufficient data for 51 out of 169 targets calls for statistical systems in the region to redouble efforts to fill data gaps*.[4] In their efforts to strengthen the production of official statistics, NSOs should explore new approaches. Data integration, which involves combining disparate data sources such as censuses, sample surveys, administrative records, and big data, is one pathway to bridging data gaps. Integration of data can shed light on obscure areas and generate insights that cannot be derived from single data sources alone.

Data integration appears to be promising; however, there are often challenges that need to be overcome to fully realise its potential, particularly in terms of institutional and technical capacities. Institutional capacity mainly refers to the availability of a) relevant data sources, b) a strong legal basis supporting the use of and access to data and metadata for statistical purposes and legislation ensuring data privacy, c) appropriate mechanisms for collaboration with data holders (either from the public or private sector), d) ICT Infrastructure, e) human resources including expertise and skills.

Technical capacities are also a challenge which includes a) skills to deal with the lack of interoperability, in terms of lack of unique identifiers, differences in concepts, classifications, coverage, data formats, reference periods, etc., b) skills to handle data quality issues, such as missing data, erroneous values, and inconsistencies within and across data sources, and c) expertise in record linkage and statistical matching. NSO staff will also require to have an in-depth understanding and knowledge of the relevant data sources, which can be obtained from appropriate metadata as well as regular interactions with data holders at the expert level. Some of the other skills that NSOs will need include soft skills, like negotiation, relationship building, and communication.

Although the capacities required for data integration appear extensive, NSOs may already have these skills and experience in data integration at the macro-level, such as in the compilation of the National Accounts, and they may be able to build on these for micro-level data integration, where capacity is limited. A regional Data Integration Capacity Assessment Survey conducted in 2020 by ESCAP revealed a significant capacity gap, with only 45 per cent of responding countries holding the required skills, as well as a huge demand for capacity building expressed by 77 per cent of responding countries, including Nepal.

4 https://www.unescap.org/kp/2023/asia-and-pacific-sdg-progress-report-2023.

Nepal project

In June 2022, ESCAP initiated a project titled "Strengthening the National Statistical Capacity to Implement Data Integration Approaches to Produce and Improve Data for Measuring SDG Indicators". This project aimed to collaborate closely with the NSO of Nepal on a practical exercise to pilot integrating data from multiple sources, including census, surveys and administrative data. The project intended to develop the capacity and provide the technical assistance and IT toolbox to implement the full process of a data integration exercise.

The first step was to identify a data integration exercise that could be completed within the project's time and resource constraints, taking into consideration the country's own priorities and available and accessible data sources. In this regard, a desk study was conducted, and views of national stakeholders were collected through an inception workshop and multiple discussions.

The defined exercise aimed to integrate relevant data sources to enable calculation of the a) proportion of the population covered by social protection floors/systems by poverty status, which will improve data for monitoring SDG Indicator 1.3.1[5] and b) produce relevant statistics to be used for validating the poverty status of households. The exercise involved the integration of four specific data sources described in Box 1.

BOX 1 – DATA SOURCES INVOLVED IN THE PILOT EXERCISE

- *Beneficiaries of Social Security Allowance 2022 (BSSA)* is a registry developed and maintained by the Department of National ID and Civil Registration, including information on program beneficiaries and social security allowances for vulnerable segments of the population.
- *Poor Households Identification Survey 2013 (PHIS)* was conducted as a large-scale survey by the Ministry of Land Management, Cooperatives and Poverty Alleviation to identify the poor for providing targeting services.
- *National Population and Housing Census 2011 (NPHC)* collected detailed information on demography, housing and asset ownership.
- *Nepal Living Standards Survey 2010/11 (NLSS)* is a multi-topic survey conducted by the NSO to produce statistics on consumption and other information on correlates of poverty.

Technical and ICT assistance was provided to facilitate accessing, preparing and linking the data sources as well as generating the specified statistics. These were followed by a training workshop in Kathmandu in February 2023[6] to enhance the technical capacity in data integration approaches, in general, and in relation to the defined exercise.

5 SDG Indicator 1.3.1: Proportion of population covered by social protection floors/systems, by sex, distinguishing children, unemployed persons, older persons, persons with disabilities, pregnant women, newborns, work-injury victims and the poor and the vulnerable.

6 https://www.unescap.org/events/2023/national-training-workshop-implementing-data-integration-approaches-official-statistics.

Enabling access to data

A key requirement for an NSO to carry out data integration is to obtain access to the required data and metadata. In Nepal, the legal basis supporting the use of and access to data and metadata for statistical purposes as well as the legislation respecting confidentiality and privacy, do exist. The NSO also had an arrangement, through official letters, with the Department of National ID and Civil Registration and the Ministry of Land Management, Cooperatives and Poverty Alleviation to access their data for statistical purposes.

However, the relevant data holders were reluctant to share their 'micro-level data' due to privacy concerns, as well as uncertainty about the feasibility of the project. To resolve the concerns and encourage collaboration, multiple meetings were organised among the NSO, data holders and the project team, in addition to the project's inception workshop. The NSO's strong commitment to protecting privacy and confidentiality in accordance with the Statistics Act[7] and the Fundamental Principles of Official Statistics [8] was underscored during these meetings. Furthermore, the project's purpose, process, methodology, and potential outputs were communicated with data holders and mutual benefits were explored.

To make the data integration exercise more manageable and to encourage data sharing from data holders, the scope of the exercise was decided to be limited to one province, Gandaki, which compared to other provinces, contained more households/persons across all data sources. Furthermore, because microdata from the most recent population census and Nepal's Living Standard Survey were unavailable, it was decided to utilise older databases. All these efforts enhanced the willingness of the data holders to share data.

Pre-processing of data sources

Integrating data across disparate sources is challenging due to the issues stemming from the need for unique identifiers, interoperability, and data quality. Hence, pre-processing is regarded as a critical step in every data integration exercise.

Pre-processing entails cleaning and standardising data to ensure that the data used for integration is accurate, complete, and consistent across different sources and it is ready for record linkage. Box 2 presents the main steps of pre-processing.

BOX 2 – DATA PRE-PROCESSING MAIN STEPS

- *Data cleaning* involves identifying and correcting errors, inconsistent values, missing values, incorrect data formats, misspelling, duplications, etc.
- *Data transformation* refers to the mapping of administrative units and/or variables into statistical units and/or variables. It also involves converting identification numbers into record identity numbers to protect privacy.
- *Data standardisation* entails ensuring consistency across data sources in terms of variable labels, values, coding, formats, spelling, etc. This step also involves identifying common variables across different sources and standardising them to a common format.

7 http://rajpatra.dop.gov.np/welcome/book/?ref=25085.

8 https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

Some of the data issues encountered in Nepal's exercise included: a) the lack of a unique identification number; b) differences in transcribing names (first, middle, and last names); c) differences in reference periods; and d) differences in classifications and categories for geographical as well as some substantive variables, such as caste and ethnic groups.

Pre-processing was conducted using a semi-automated approach, as some issues required manual resolution. The issues related to the translation of names from Nepali to English were resolved by thoroughly examining the data using Nepali naming conventions. There were also problems with the spelling of names, which were corrected manually. Standardised coding was developed to ensure consistent attribute labels and categories across all databases. The difference in reference periods was addressed by adjusting for the difference in ages.

Due to the lack of a unique identifier across data sources, it was decided to use probabilistic matching based on names, location, and demographic variables, as appropriate.

Linkage of data sources and the potential output

Record linkage refers to the identification and combination of records corresponding to the same individuals across two or more data sources. There are two main approaches to record linkage, deterministic and probabilistic matching.

Deterministic matching involves comparing records in two datasets to find an exact match based on a unique identifier. In contrast, probabilistic matching uses statistical methods to determine the likelihood that two records refer to the same entity. While deterministic matching is straightforward, it can result in missed matches if the criteria are too strict or false matches if the criteria are too lenient. Probabilistic matching, although more complex, can result in more accurate matches by assigning weights to different matching criteria based on their importance and calculating a probability score for each potential match. Overall, the choice of approach depends on the specific requirements of the application, including the quality of the data, the level of accuracy required, and the resources available.

In Nepal’s exercise, due to the lack of a unique identification number across data sources, deterministic matching was not possible, so probabilistic matching methods had to be applied, using common variables. To facilitate the record linkage, an IT toolbox was developed, which provided the user with the required features to explore data sources, set the rules and run the linkage process.

The toolbox also generated specified statistics. In this exercise, the toolbox provided cross-tabulations of variables from two different data sources. For instance, a table of the type of social security allowance from BSSA versus poverty status from PHIS was provided, which enabled the calculation of the proportion of the population covered by the social protection system based on poverty status. Also, wealth quintiles from the NPHC were cross-tabulated versus the poverty status from PHIS. This table enabled estimating the percent of those households in the bottom quintile of the wealth index that had been identified as poor in PHIS, allowing validation of the poverty status calculated in PHIS.

Conclusion, limitations, and the way forward

The exercise created a prototype for future data integration efforts in Nepal. It demonstrated the integration of multiple data sources, including the population census, sample surveys and administrative data. The exercise covered all steps involved in a data integration activity, from setting the objective for data integration to accessing data sources, pre-processing data, linking records and finally generating new statistics within a single province, Gandaki. While the exercise was successful in many ways, some limitations should be acknowledged, such as that more recent data would have been more useful and policy-relevant.

The capacity that has been built in NSO and project stakeholders enables them to expand the scope of the pilot exercise to the entire country, as well as to initiate other similar data integration projects. Although the current exercise focused on integrating data from census, survey, and administrative data sources, the NSO may also wish to explore the use of new data sources, such as geospatial information and/or big data.

NSOs interested in integrating data from multiple sources should consider developing or strengthening mechanisms for collaboration and coordination with other government agencies, non-governmental organisations (NGOs), the private sector, and academia. These mechanisms can take the form of formal agreements with relevant parties, such as a Memorandum of Understanding (MoU), [9] or they can refer to activities that strengthen relationships at the senior management and expert levels, such as forums, regular meetings, or joint initiatives.

The NSOs should also take further actions to advance the use of administrative data in producing official statistics. With many countries opting to further benefit from available administrative data, a wealth of resources is now available. [10]

Further capacity-building is required both for NSOs and their national stakeholders. Additionally, awareness-raising activities are necessary to communicate to policymakers, the media, and civil society the benefits of using and integrating available data sources for producing official statistics. Building a culture of data use would ensure that integrated data is effectively utilised in decision-making, leading to more informed and effective policy outcomes.

At the regional level, it is recommended to explore mechanisms for expanding the technical support on data integration activities. This may include developing e-Learning courses and conducting regional training workshops on data integration, replicating the project in other countries, expanding the current toolbox into a general toolbox applicable to various data structures, and/or encouraging more developed NSOs to provide technical assistance to NSOs in need of assistance in the region.

This is unquestionable that NSOs can no longer meet all data demands in isolation. The data landscape is evolving, and to catch up with this changing world, NSOs should step out of their comfort zone and explore new horizons. Benefitting from available knowledge and experiences and enhancing statistical capacity makes the journey smoother.

9 See the draft template developed by United Nations Statistics Division at https://unstats.un.org/capacity-development/admin-data/docs/mou-guide-and-template.pdf.

10 See the inventory of the Collaborative on the Use of Administrative Data for Statistics at https://unstats.un.org/capacity-development/admin-data/#.

Get connected. Follow us.

www.unescap.org

instagram.com/unitednationsescap

facebook.com/unescap

youtube.com/unescap

twitter.com /unescap

linkedin.com/company/united-nations-escap

Previous issues of Stats Brief: http://www.unescap.org/resource-series/stats-brief