Asia-Pacific Guidelines to Data Integration for Official Statistics

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

*The shaded areas of the map indicate ESCAP members and associate members**

The Economic and Social Commission for Asia and the Pacific (ESCAP) is the most inclusive intergovernmental platform in the Asia-Pacific region. The Commission promotes cooperation among its 53 member States and 9 associate members in pursuit of solutions to sustainable development challenges. ESCAP is one of the five regional commissions of the United Nations.

The ESCAP secretariat supports inclusive, resilient and sustainable development in the region by generating action-oriented knowledge, and by providing technical assistance and capacity-building services in support of national development objectives, regional agreements and the implementation of the 2030 Agenda for Sustainable Development.

**The designations employed and the presentation of material on this map 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.*

PREFACE

Work on official statistics in Asia and the Pacific is performed in close collaboration with the 53 member States and 9 associate members of the Economic and Social Commission for Asia and the Pacific (ESCAP) through consensus building, knowledge and analytical products, and technical cooperation. Consensus building is primarily achieved through the ESCAP Committee on Statistics, a biennial meeting of heads of national statistical offices (chief statisticians). Under the auspices of the Committee on Statistics, regional mechanisms have been established to present the Asia-Pacific statistical community with opportunities to meet and exchange ideas, and to develop standards and good practices across all fields of official statistics.

In 2018, the Committee on Statistics, at its sixth session, held in Bangkok from 16 to 19 October 2018, noted the importance of integrated statistics and supported a proposal by the secretariat[1] to set up communities of practice as online modalities to further the work around integrated statistics.[2] Later, the Bureau of the Committee on Statistics, in its meeting held in December 2018, agreed on establishing a data integration community of practice as an immediate priority.

In 2019, the Regional Steering Group on Population and Social Statistics, at its second meeting, held in Bangkok from 17 to 19 July 2019,[3] guided by the Committee's decision, and inspired by a Stats Brief by the secretariat,[4] focused on issues related to integrated statistics. During this meeting, members decided that the development of guidelines on data integration is a top priority. They also supported trialling the application of a community of practice, as an online learning platform for enhancing the sharing of knowledge and experience relating to data integration.

In late April 2020, as per the decisions made by the Committee on Statistics and the Regional Steering Group on Population and Social Statistics, the secretariat launched the Data Integration Community of Practice (DI-CoP), mainly for practitioners in the Asia-Pacific statistical community that have expertise or interest in learning about data integration. The first priority of the Community of Practice was to focus on the development of guidelines on data integration relevant to the Asia-Pacific region, building on a guide developed by the Economic Commission for Europe (ECE), through the High-Level Group for the Modernization of Official Statistics.[5]

---

[1] See ESCAP/CST/2018/1, p. 4.
[2] See ESCAP/CST/2018/6.
[3] See https://www.unescap.org/events/second-meeting-regional-steering-group-population-and-social-statistics.
[4] Economic and Social Commission for Asia and the Pacific, "Integrated statistics, a journey worthwhile", Stats Brief, Issue no. 19 (July 2019). Available at https://www.unescap.org/sites/default/files/Stats_Brief_Issue19_Jul2019_Integrated_Statistics.pdf.
[5] United Nations, Economic Commission for Europe, "A guide to data integration for official statistics", version 2. Available at https://statswiki.unece.org/display/DI/Guide+to+Data+Integration+for+Official+Statistics.

The Asia-Pacific Guidelines to Data Integration for Official Statistics was developed using contributions from members of DI-CoP. It was also informed by the results of the Data Integration Capacity Assessment Survey (DI-CAS)[6] conducted in 2020 (a summary of the results is provided in the annex) and inputs from participants in regional workshops on implementing data integration in Asia and the Pacific.[7]

The Asia-Pacific Guidelines to Data Integration for Official Statistics gives practical advice and information to advance data integration activities by statistical organizations. The guidelines are relevant to managers, statisticians, methodologists, information and communications technology (ICT) professionals and other staff members of statistical organizations who use or aspire to use data integration in the production of official statistics. It provides information about issues that statistical organizations have or should consider in work related to data integration.

[6] The Data Integration Capacity Assessment Survey (DI-CAS) was adapted from ECE survey in 2017 (https://statswiki.unece.org/display/DI/2017+Data+Integration+Survey). DI-CAS full results are only available to members of DI-CoP. To access full results, please join the DI-CoP at https://stat-confluence.escap.un.org/pages/viewpage.action?spaceKey=DICP&title=Data+Integration+Community+of+Practice.

[7] See https://www.unescap.org/events/regional-workshops-implementing-data-integration-asia-and-pacific-round-1.

ACKNOWLEDGEMENTS

The Asia-Pacific Guidelines to Data Integration for Official Statistics is the result of collective efforts of the members of the Data Integration Community of Practice (DI-CoP) to adapt the guidelines developed by ECE for the Asia-Pacific region.

The first and foremost gratitude is expressed to the ECE Statistics Division, particularly Lidia Bratanova, Steven Vale and Tiina Luige for their generous support.

The Guidelines would not have been possible without the invaluable contributions and support from members of DI-CoP, in particular Karine Kuyumjyan (Armenia), Siu-Ming Tam (Australia), Soheil Rastan (Canada), Rebecca Wai Fun Siu (Hong Kong, China), Ashutosh Ojha (India), Yuni Arti (Indonesia), Sayed Mohammd Hosseini, Zahra Rezaei Ghahroodi, Mohaddesseh Safakish and Ashkan Shabbak (Islamic Republic of Iran), Nuramalina Abdullah (Malaysia), Will Bell, Akib Mohammad, Aravindh Rajendran and Gus Segura (New Zealand), Aliimuamua Malaefono Taua (Samoa), Aycan Özek (Turkey), Arturo Martinez (Asian Development Bank), Shiomi Yumi (Asian Disaster Reduction Center), Irina Bernal, Jenine Borowik, Arman Bidarbakhtnia, Jessica Gardner, Gemma Van Halderen, Ayodele Marshall, Petra Nahmias, Alick Nyasulu, Sharita Serrao, Dayyan Shayani and Afsaneh Yazdani (ESCAP), Alison Culpin, Nilima Lal and Gloria Mathenge (Pacific Community), Jayachandran Vasudevan (UNICEF) and Jonathan Gessendorfern (United Nations Department of Economic and Social Affairs Statistics Division).

The Guidelines was informed by the responses to the Data Integration Capacity Assessment Survey (DI-CAS), provided by the national statistical offices of Armenia; Australia; Bangladesh; Bhutan; Brunei Darussalam; Cambodia; Fiji; Georgia; Hong Kong, China; India; Indonesia; Islamic Republic of Iran; Japan; Kazakhstan; Lao People's Democratic Republic; Malaysia; Marshall Islands; Mongolia; Myanmar; Nepal; New Zealand; Pakistan; Philippines; Republic of Korea; Russian Federation; Samoa; Singapore; Thailand; and Turkey, as well as the National Organization for Civil Registration of the Islamic Republic of Iran and the Ministry of National Health Services, Regulations and Coordination of Pakistan.

The Guidelines also benefited from insights from participants of the regional workshops on implementing data integration in Asia and the Pacific, organized in November and December 2020.

This work also received immense support from Krisana Boonpriroje, Nasikarn Nitiprapathananun and Panita Rattanakittiaporn and from the ESCAP Information Management, Communications and Technology Section, specifically from Elom Etse.

Thanks also are extended to Alan Cooper for reviewing and editing to ensure clarity of language, and Nandini Khandekar for the formatting and graphic design of the Guidelines.

The Economic and Social Commission for Asia and Pacific acknowledges financial support attained through the United Nations Development Account (10th tranche) project entitled “Data and Statistics”.

TABLE OF CONTENTS

Preface iii

Introduction 1

What is data integration? 5

Planning for data integration 9

3.1. Access to data 12

3.2. Partnerships 14

3.3. Skills 16

Data considerations 18

4.1. Concepts in the data 21

4.2. Identifiers 23

4.3. Privacy and confidentiality 24

Quality 27

Methods and tools 34

6.1. Record linkage 35

6.2. Statistical matching 38

6.3. Software tools 40

6.4. Other methodological considerations 42

Communicating integrated data 43

7.1. Key audiences 44

7.2. What is different about communicating integrated data? 46

7.3. Examples of communicating data integration 47

Types of data integration 49

8.1. Traditional data sources (surveys, censuses, administrative sources) 50

8.1.1. Administrative sources 50

8.1.2. Surveys and censuses 54

8.2. New sources of data (geospatial, big data) 58

8.2.1. Geospatial information 58

8.2.2. Big data 63

8.3. Validating and improving official statistics 69

Final comments 71

Annex 72

ABBREVIATIONS AND ACRONYMS

ADB	Asian Development Bank
CPI	consumer price index
DI-CAS	Data Integration Capacity Assessment Survey
DI-CoP	Data Integration Community of Practice
ECE	Economic Commission for Europe
ESCAP	Economic and Social Commission for Asia and the Pacific
GIS	Geographic Information System
GSBPM	Generic Statistical Business Process Model
GSGF	Global Statistical Geospatial Framework
ICT	information and communications technology
IT	information technology
PII	personally identifiable information
SAE	small area estimation
SDG	Sustainable Development Goals
UN-GGIM-AP	United Nations Global Geospatial Information Management for Asia and the Pacific
UNICEF	United Nations Children's Fund

Chapter 1

INTRODUCTION

1. Increasingly, new data sources are becoming available to statistical organizations. This comes at a time when modern technologies are available to support data integration, which makes it possible to produce timelier, more disaggregated statistics at higher frequencies in a more cost-effective manner, as compared with traditional approaches.

2. Data integration can be used to provide new official statistics, address new or unmet data needs, reduce response burden, overcome the effects of declining response rates, and deal with quality and bias issues in surveys. In order to meet national and international reporting commitments under the 2030 Agenda for Sustainable Development, greater focus is being placed on data integration around the world.

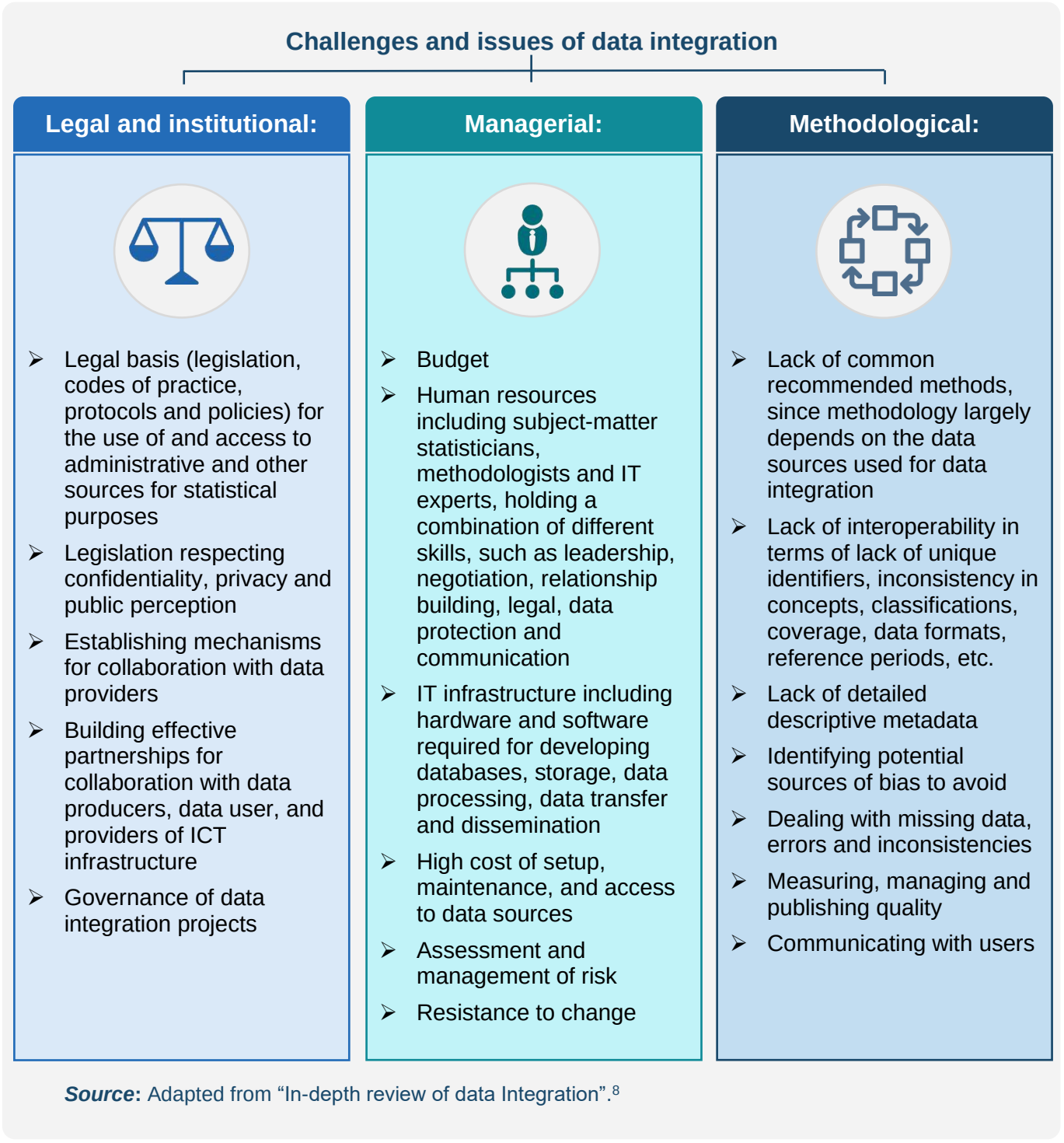
3. Statistical organizations are being challenged by the need to integrate diverse sets of inconsistent data and produce stable outputs with sometimes unstable, ever-changing inputs. Instead of trying to produce the best possible statistics from a single data source, it is necessary to find the best combination of sources to deliver the indicators or statistics that most efficiently satisfy the users' needs.

4. Some potential challenges related to data integration are the following:

    - Enabling an institutional environment that includes legislation, coordination and collaboration mechanisms, partnerships with data providers, commercial companies and academia, as well as set-up and ongoing costs;
    - Dealing with diverse forms of data integration;
    - Developing new skills, methods, and information technology (IT) approaches;
    - Designing new concepts or aligning existing statistical concepts with the concepts in new data sources;
    - Establishing or selecting appropriate ICT infrastructure, storage, maintenance and transfer protocols;
    - Quality assurance for statistics produced by integrating different data sources;
    - Developing data integration projects from the stage of being research projects to repeatable, and reliable production of statistics;
    - Establishing governance for data integration projects, within and across organizations;
    - Managing public perception and communication;
    - Addressing issues around data confidentiality and privacy when dealing with personal data;
    - Defining ownership of the statistics produced where multiple data providers are involved;
    - Avoiding duplication of efforts across and within countries and organizations and using the collective experience of official statistics communities.

5. In figure 1, the different challenges and issues that come with data integration are categorized.

Figure 1. Challenges and issues of data integration



***Source*:** Adapted from "In-depth review of data Integration".[8]

[8] United Nations, Economic Commission for Europe, "In-depth review of data Integration", prepared by the secretariat and the participants and project leader of the HLG-MOS 2016 Data Integration Project for Conference of the European Statisticians, Geneva, 14–15 February 2017. (ECE/CES/BUR/2017/FEB/2). Available at https://unece.org/DAM/stats/documents/ece/ces/bur/2017/February/02_in-depth_review_data_integration_final.pdf.

6. In 2020, ESCAP conducted DI-CAS, which was adapted from the ECE survey. The results from DI-CAS suggest that the most common barrier to data integration faced by the responding organizations is “maintaining access to data sources”, followed by “public acceptance and trust issues”, “ICT issues”, “lack of supporting legislation”, “not having access to data sources” and “lack of relevant methodologies”.

7. To tackle these challenges and issues, this guide provides information on planning for data integration, issues related to data, data integration methods and tools, and presents relevant examples from the Asia-Pacific region.

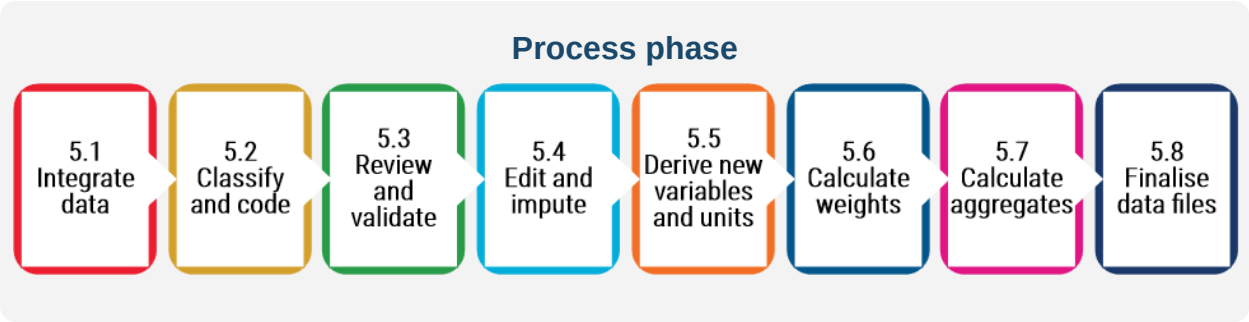
Chapter 2

WHAT IS DATA INTEGRATION?

8. To extend guidance for statisticians on how data integration activities fit into the statistical business process, it is important to define the term “data integration”.

9. Based on the Generic Statistical Business Process Model (GSBPM)[9] data integration is defined under the “Process Phase”, which is a phase that describes the processing of input data and their preparation for analysis. This phase is comprised of subprocesses that integrate, classify, check, clean and transform input data, so that they can be analysed and disseminated as statistical outputs. Data integration may occur at any point in this phase, before or after any of the other subprocesses. Several instances of data integration may also occur in any statistical business process. Following integration, depending on data protection requirements, data may be de-identified, namely stripped of identifiers, such as name and address, to help to protect confidentiality.

Figure 2. Generic Statistical Business Process Model (GSBPM)



10. Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information. Based on GSBPM, input data can be from a combination of external or internal sources, and a variety of collection instruments, including extracts of administrative and other non-statistical data sources, such as big data. Administrative data or other non-statistical sources of data can be used as substitutes for all or some of the variables directly collected from surveys. This process also entails harmonizing or creating new figures that agree between sources of data. Data integration can include the following:

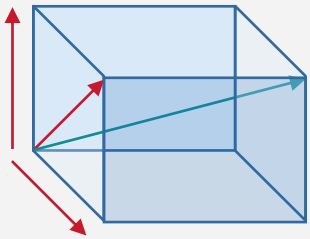
    - Combining data from multiple sources, as part of the creation of integrated statistics, such as national accounts;
    - Combining geospatial data and statistical data or other non-statistical data;
    - Data pooling, with the aim of increasing the effective number of observations of some phenomena;
    - Matching or record linkage routines, with the objective to link micro or macro data from different sources;
    - Data fusion – integration followed by reduction or replacement;
    - Prioritizing, when two or more sources contain data for the same variable, with potentially different values.

11. One conceptual model for data integration is a cube (figure 3) that holds multilayered data in different standard classifications and aggregations, such as survey data, administrative records, big data or an estimate. While traditional data integrations deal with

[9] See https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

the red lines (face horizontal integration, face vertical integration, face diagonal integration), advanced data integrations deal with space diagonal integration (the blue-like line).

Figure 3. Conceptual model for data integration



12. Data integration is about bringing two or more datasets together to produce value for data users, often focusing on policymakers. Accordingly, official statisticians should first identify the data gaps and official statistics required to inform the policy priorities and then consider data integration as another tool available to bridge the gaps. Data integration can help policymakers and researchers gain a much better understanding of local, national, regional and international circumstances of people, communities, industry, and the economy. This can help to improve the development and delivery of governmental services in such areas as health, education, infrastructure maintenance and development and other community services. One very good example is in Australia where the Australian Bureau of Statistics integrates data to enhance data availability to inform the country's important decisions.[10]

13. The result of a data integration activity is always an integrated dataset. Some examples of data integration are the following:

    - Datasets combined to produce a sampling frame for a survey, such as building statistical business registers (see *User Guide for ADB Statistical Business Register*).[11]
    - Several sources integrated into one dataset to provide microdata files to researchers for statistical purposes, such as the Australian Bureau of Statistics Multi-Agency Data Integration Project (MADIP) [12] and the Statistics New Zealand Integrated Data Infrastructure (IDI).[13]
    - An integrated dataset that serves as an input to produce official statistics. For example: ABS – Improving the Labour Account.[14]

[10] For more information, please refer to https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Data+Integration.
[11] Asia Development Bank, User Guide for ADB Statistical Register (Manila, ADB, December 2018). Available at https://www.adb.org/publications/adb-statistical-business-register-user-guide.
[12] See https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Multi-Agency%20Data%20Integration%20Project%20(MADIP).
[13] See https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/.
[14] Australian Bureau of Statistics, "Labour account Australia methodology" (September 2020). Available at https://www.abs.gov.au/methodologies/labour-account-australia-methodology/sep-2020.

- Data from several subject-matter domains combined into one dataset used as the basis for production of statistics (such as national accounts).
- Datasets from different subject-matter domains compared to check the quality and the validity of information produced (macro-validation).
- A dataset integrated for the purposes of micro-validation when some rules are defined to check the validity of the data in one dataset compared to another one.
- Missing values imputed in a dataset using another dataset as the source for imputation.
- A statistical model developed and produced using different sources to produce model-based information. For example, the Japan's experience in forecasting rice yield using satellite data.[15]

More examples are available on the DI-CoP platform as resources and case studies.

14. In the ESCAP 2020 DI-CAS, statistical organizations were asked about their data integration experiences and practices. The results showed that the most important or prominent ways the responding organizations were using data integration or alternate data sources were to supplement surveys or traditional censuses, such as to complement survey for a subpopulation, or for a set of variables, or to conduct register-based censuses. Data integration was also commonly used as a source for building sampling frames and for data validation and/or imputation and less prominently to replace sample surveys or traditional censuses. The survey results also reflect that data integration was being used most for the ongoing production of social and economic statistics and much less for environmental statistics. There was a high level of experimentation and research in line with the aspiration to increase the use of data integration, particularly to produce or improve Sustainable Development Goals indicators.

[15] Masahiro Hosako, "Forecast on rice yield using satellite data", presentation. Available only to DI-CoP members at https://stat-confluence.escap.un.org/download/attachments/1409148/Forecast%20on%20rice%20yield%20using%20satellite%20data%20%28Hosaka%20Masahiro%2C%20Japan%29.pdf?version=1&modificationDate=1611628404524&api=v2.

Chapter 3

PLANNING FOR DATA INTEGRATION

15. Data integration requires a range of technical and institutional capacities. An enabling institutional environment comprises development of appropriate legislative frameworks, building formal and informal coordination and collaboration mechanisms and partnerships with data providers, commercial companies and academia, among others.

16. Data integration for producing official statistics is often led by national statistical offices in collaboration with other organizations and data holders inside or outside the national statistical system. National statistical offices tend to have in place some of the prerequisites, such as legislation, trust of citizens, data skills and confidentiality tools. Moreover, national statistical offices are the custodian of the population census dataset, one of the most (if not the most) powerful national datasets.

17. Statistics New Zealand considers the steps shown in figure 4 for each data integration project. Each step is fundamental, and not one can be omitted.

Figure 4. Key steps in a data integration project

***Source*:** *Statistics New Zealand, Data Integration Manual: 2nd Edition* (Wellington, New Zealand, Statistics New Zealand, 2013).

18. Provisions and arrangements may be different between centralized and decentralized statistical systems. However, some of the most common activities when planning for data integration are the use of cooperation agreements for transferring data, preparation of legal documents for establishing or maintaining use of the data, developing long-term partnerships (formal or informal), which consist of two or more organizations using the same data, and consideration of guidelines for accessing data among organizations (either in the public and private sectors). To these activities, planning to develop and implement standard concepts and classifications at the national level can also be added. All of these activities can support ongoing collaboration between parties. A favourable scenario would include a national or central and official statistical agency with the legal and moral authority to organize an integrated data framework by liaising with multiple organizations across a country or territory. In contrast, it is difficult to organize an integrated data framework if State and/or territorial authorities overlap or compete with each other.

19. Cooperation agreements are often signed to do the following: (a) divide the tasks among the parties of the agreement; (b) define the rules and conditions of transferring data, such as timeliness, metadata and reidentifying (or not) individuals for personal data, and (c) define technical implementation (for example where the data will be stored, the use of shared platforms, and/or provision of coding or auto-coding technology to the parties). Data-sharing agreements are intended to systematize and mainstream data exchange partnerships among different data holders. A comprehensive discussion of this can be found in a report published in June 2020 by Hayden Dahmm of the UN Sustainable Development Solutions Network for the Thematic Research Network on Data and Statistics entitled “Laying the foundation for effective partnerships: an examination of data sharing agreements”.[16] The report also includes a discussion on other issues to be considered, including data use, access, handling of breaches, proprietary issues, publicization of the analysis and deletion of data upon termination of the agreement.

20. Results of the ESCAP 2020 survey (DI-CAS) show that many statistical organizations have developed or are developing international, national and/or organization-wide strategies for data integration. Many countries already have in place laws that give their respective national statistical office full or partial access to government data, and some even have in place a law that supports access to private sector data. Nevertheless, privacy issues remain a serious barrier to integrating data in many instances. The results of the survey also show that collaboration and agreements with other organizations are common practices.

[16] Hayden Dahmm, “Laying the foundation for effective partnerships: an examination of data sharing agreements” (12 June 2020). Available at https://www.sdsntrends.org/research/dsainsightsreport.

3.1. ACCESS TO DATA

21. A key requirement for data integration is to have access to the desired data (micro and macro) and metadata. In most countries, data from either the government or private sector can be accessed freely by the national statistical office. In some countries and cases, data are only available upon payment, while in others, the data exist but cannot be accessed. Privacy and confidentiality issues sometimes hinder access to data by national statistical offices.

22. A legal basis is often important to provide national statistical offices with access to data for statistical purposes. A sound approach is to ensure national statistics legislation supports use of already existing data sources, such as administrative data and big data, rather than recollecting data. The use of administrative sources for statistical purposes is often stated in a statistical act, which is needed to consider various statistical, methodological, legal, and ethical issues, and include new data sources as well.

23. In addition to the legal basis, it is important to build other mechanisms that facilitate close collaboration and exchange of data and metadata among stakeholders.

24. Practical work to develop common approaches starts with data. Some types of publicly available data can be used without much difficulty, such as Integrated Public Use Microdata Series (IPUMS) data,[17] some web-scraped data or social media sources, government-owned open data or public statistical outputs, if the right access, tools, skills and techniques are on hand.

25. Using an experimental dataset makes the initial phase in the implementation of new collaborative data integration projects easier. It is also possible to systematically de-identify real data (from surveys, censuses, government registrations or privately held big data sources). The creation and documentation of a set of synthetic datasets allows different organizations to collaborate on developing common methods, removing issues of confidentiality and encouraging use of the same data formats. As suitable methods, processes and tools are developed in a collaborative way, they can be moved to the secure environments of individual organizations for further testing on real data. Although this may be done within a particular country, the potential to bring some of these data providers and a group of statistical organizations together to explore mutual benefits and potentially develop global or regional agreements for data supply should also be explored.

26. Enhancing the quality of the consumer price index (CPI) in terms of coverage and real-time quantity is a common challenge faced by many countries. Data integration can help in this effort. Several data providers operate globally or in many countries, opening up the opportunity to develop a common approach that can be used in multiple countries. Some organizations have significant data holdings. Among them are IRi, Worlddata.Info and Numbeo.[18]

27. Many issues require early consideration when establishing approaches to obtain and

[17] For more information about IPUMS, see https://ipums.org/.
[18] For more information about these organizations, please refer to their respective websites: IRi: https://www.iriworldwide.com/en-au/company/about-us; WorldData.info: https://www.worlddata.info/cost-of-living.php; Numbeo: https://www.numbeo.com/cost-of-living/.

secure access to data and negotiate with data providers. Among them are the following:

- Some countries may not have access to the same data sources or formats; levels of detail or aggregation may vary across countries;
- Data protection policies in some countries may restrict access to certain types of data;
- Use of new data sources, such as satellite data and big data, may impose an extra financial burden, in terms of payments to data providers or the need for access to additional ICT capacity;
- Access to application programming interfaces may be restricted under terms and conditions;
- Commercial companies may not be interested in partnering with statistical organizations unless a compelling case is made;
- There is potential for interrupted supply, while producing official statistics requires secured access to data;
- Delayed receipt of data may reduce the representativeness of the data for reference periods;
- Maintaining the social licence of the data integration initiative, associated research projects and data suppliers to minimize objections of stakeholders and to support ongoing access to the data. (Further information about social licence is given in section 4.3);
- Sometimes data are only accessible in hard copy rather than in digital form, which requires migration from paper-based to computerized databases;
- Data sharing among different government organizations may be challenging, especially in situations in which each governmental organization owns and manages its own data;
- Managing data that are scattered across several organizations, and at different levels of government may be cumbersome;
- Use of a common area for the lodgement and storage of data (not necessarily in statistical organizations) may be helpful.

3.2. PARTNERSHIPS

28. The importance of establishing and maintaining effective partnerships for data integration should be recognized. Partnerships can be formed with the following entities or groups:
    - Other statistical organizations to collaborate and share experiences, approaches, and standards;
    - Public sector data providers;
    - Private sector data providers, which often have different behaviours and requirements than public sector data providers;
    - Institutions within countries, such as tax offices, employment office and registries – together deciding on the methodology, concepts and classifications;
    - Technology organizations and ICT infrastructure providers;
    - Research initiatives and academia;
    - Political leaders;
    - Users of statistics and citizens.
29. Among the many actions or initiatives that encourage effective partnerships are the following:
    - Establishing personal and friendly connections between organizations and relationship management;
    - Providing feedback on the data regarding their usefulness for official statistics needs;
    - Promoting the goals of the data integration project to the providers and jointly clarifying mutual benefits and opportunities (“what is in it for each organization”);
    - Encouraging evidence-based or data-driven partnerships that increase the confidence of stakeholders for effective partnership (as best evidence complements the decision-making);
    - Facilitating data exchange among public organizations – including exploring opportunities for organizations to give microdata in return for providing value added statistics and data;
    - Understanding and managing barriers, such as costs, capacity and risks;
    - Establishing formal agreements;
    - Encouraging the national statistical office to take a leadership role in the development of relevant laws, agreements and procedures for use by partners and stakeholders;
    - Setting a service framework with dedicated working teams to manage relationships with suppliers, researchers, and the general public, as well as to support all integrated data operations and safeguards;
    - Depending on the need within countries, encouraging the national statistical office to provide systems or software, standards, classifications, coding tools, expert resources and/or training on data integration for other government organizations;

- Engaging with the country's leaders to educate and seek their support regarding the benefits of data integration, including the development of national capacity for data integration and development of statistical laws;
- Engaging with the ministries and organizations involved in particular statistical domains to understand their drivers and approaches and agree on standards and metadata (examples of domains may be education, health and financial);
- Engaging with the public to build trust through the media and relevant examples;
- Establishing specific joint projects related to the production of statistics which meet data and statistics demands for one or more of the Sustainable Development Goals. (examples of these can be found in the following reports, publications or presentations: Mapping poverty through data integration and artificial intelligence;[19] Poverty mapping and disaggregated estimates using Small Area Estimation in ECLAC;[20] Using and integrating non-traditional data sources in urban monitoring;[21] and Data protection and integration: improving availability of data on vulnerable children (SDG 1.3.1)[22]).

30. Further suggestions about forming effective partnerships and communicating well with partners and others are provided in chapter 7. Some issues are best tackled through individual partnerships, while others may benefit from country-level, multi-country, regional or even global approaches. Even in cases in which countries have their own variations on issues, use of a regional starting point, such as this guide, and regional meetings or workshops, can provide ideas for addressing the issues.

[19] Arturo M. Martinez Jr., “Mapping poverty through data integration and artificial intelligence”, presentation (November 2020). Available at https://www.unescap.org/sites/default/files/Session4_ADB_mapping_poverty_DI-CoP_WS.pdf.

[20] Andrés Gutiérrez, “Poverty mapping and disaggregated estimates using small area estimation in ECLAC”, presentation (November 2020). Available at https://www.unescap.org/sites/default/files/Session4_ECLAC_poverty_mapping_DI-CoP_WS.pdf.

[21] Robert Ndugwa, “Using and integrating non-traditional data sources in urban monitoring”, presentation (November 2020). Available at https://www.unescap.org/sites/default/files/Session4_UNHabitat_Urban_monitoring_DI-CoP_WS.pdf.

[22] Pedro Freire, “Data protection and integration: improving availability of data on vulnerable children” (SDG 1.3.1), presentation (November 2020). Available at https://www.unescap.org/sites/default/files/Session4_UNICEF_data_on_vulnerable_children_DI-CoP_WS.pdf.

3.3. SKILLS

31. Statistical organizations should plan for their staff members to develop new skill sets that enable them to harness new technologies, design and apply complicated statistical methodologies, understand complex legal and policy issues, and effectively communicate and negotiate. In addition to comprehensive capacity development programmes, part of technical capacity-building can be fulfilled through regular and structured user-producer dialogues. Exchange of knowledge and experience among national statistical system stakeholders through regular formal and informal meetings at different organizational levels would also be beneficial.

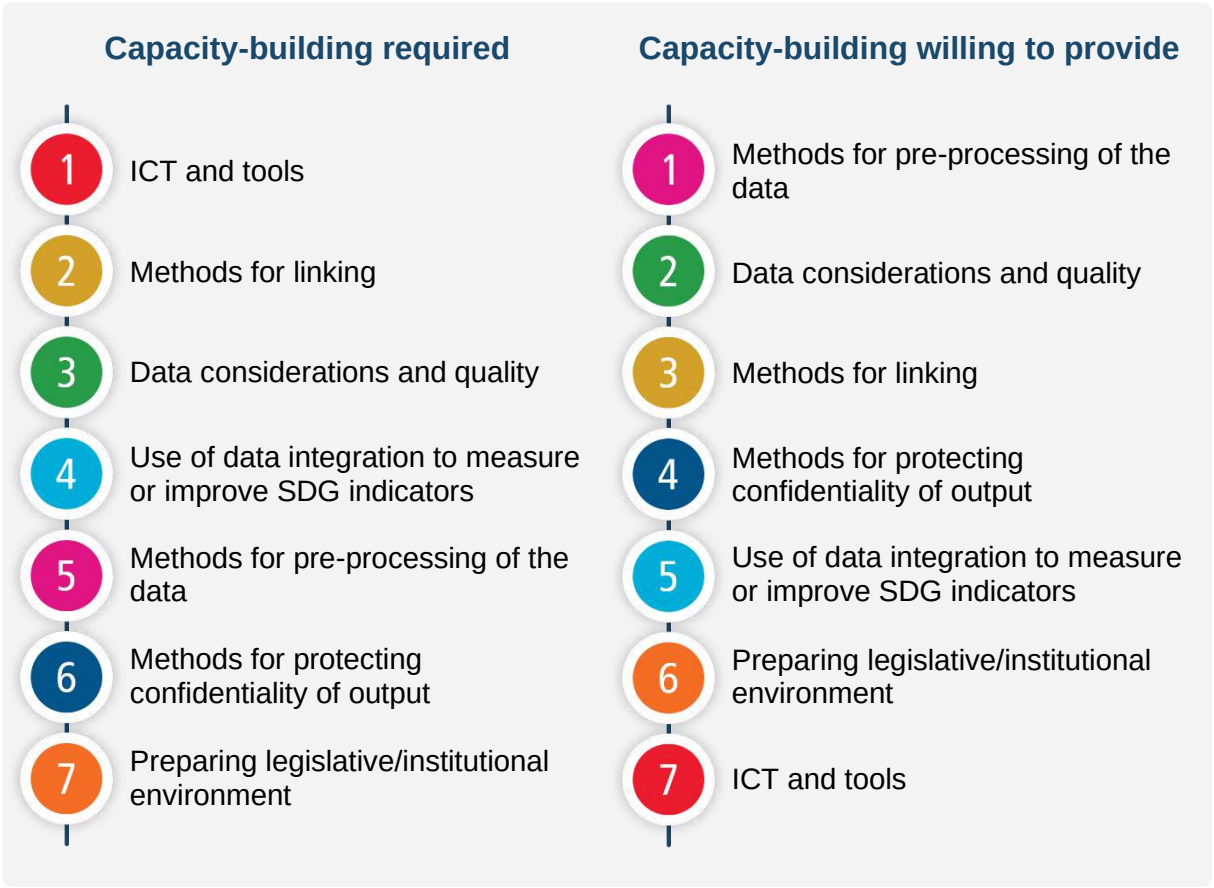
32. Some of the essential skills needed for integrating data are the following:

- Leadership and negotiation skills for participating in policy development and in discussions with data providers;
- Legal skills related to the legal basis for obtaining data, data protection and cooperation agreements with data providers (either in the public or private sectors);
- Subject-matter knowledge and expertise in understanding data needs, data content, statistical processes and dissemination methods;
- Methodological skills related to all statistical processes, such as preparation of sampling frame and selection of observation units, data linkage and matching, weighting, time series analysis and seasonal adjustment.
- Social studies and privacy skills to be able to detect risks associated with integrating data that contains personal information and to be aware of issues that can affect different groups and segments of the population.
- Programming, software and database skills for construction of microdata databases and for establishing and maintaining generic and non-generic process programmes, such as for data editing and imputations, validation, aggregation and tabulation, micro and macro data analysis, data protection, encryption, integration processing and managing access permissions;
- Business architecture expertise;
- Understanding of data models, data mapping documents, ETL (extract, transform and load) design and coding;
- Ability to communicate effectively with data providers, data users and providers of ICT infrastructures.

33. In the 2020 ESCAP survey (DI-CAS), organizations were queried about their skills level and interest in obtaining or providing skills development related to data integration. The results showed that less than half of the responding organizations held or had access to the required skills to undertake data integration activities. Around one third of the responding organizations had developed specific training in data integration or in some other form of capacity-building in recent years. This presents an opportunity to tap into existing training programmes to support capacity development in other countries. Almost all responding organizations expressed interest in receiving training or capacity development and approximately half of the respondents indicated interest in providing training or capacity development to others. The most common forms of training or capacity development needed or can be provided are indicated in figure 5. It is interesting that ICT

and tools ranks highest in terms of needs (left column) but had the lowest ranking among those willing to provide support. There was a closer alignment between needs and offerings for “methods for linking” and “data considerations and quality”.

Figure 5. Common forms of training or capacity development needed or can be provided by responding organizations in the 2020 ESCAP survey (DI-CAS)



Chapter 4

DATA CONSIDERATIONS

34. Many issues need to be discussed when considering the datasets to be integrated, such as conceptual alignment of the datasets, identifiers and privacy concerns. These can be categorized under interoperability, which is a term covering broad aspects of data and metadata exchange.

35. Data interoperability addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data (Data Interoperability Standards Consortium).[23] Interoperability is the ability to join-up and merge data without losing meaning (JUDS 2017).[24] In practice, data are considered interoperable when they can be easily reused and processed in different applications, allowing different information systems to work together (Data Interoperability: A Practitioner's Guide to Joining Up Data in the Development Sector).[25]

36. The Data Commons Framework, which was devised by Elena Goldstein, Urs Gasser and Ryan Budish in 2018, splits the concept of interoperability into a number of narrow and broad layers that relate to standardization and semantics. These layers can help in the development of projects, plans, and road maps to better understand interoperability needs at various points. Figure 6 provides a summary of the layers.[26]

37. The distinction between "availability" and "accessibility" of data should be clearly made and communicated across a statistical system. Often, available data are not accessible (and accordingly not usable) for various reasons, such as lack of availability at the desired level of disaggregation, absence of relevant variables, lack of sufficient documentation (metadata) required for understanding and using data, lack of open data strategy by data custodian, inappropriate data structure and format, and in general, lack of interoperability.

38. Certain technologies, protocols and standards are required to be adopted by data custodians to establish interoperability. International standards, such as the Statistical Data and Metadata eXchange (SDMX),[27] are intended to provide standard models that make datasets interoperable and make data integration more feasible. Use of application programming interfaces by various data custodians is another technology that provides machine-to-machine access to data by different users and enhances data interoperability and, as a consequence, integration.

[23] See https://datainteroperability.org/.
[24] Liz Steele and Tom Orrell, "The frontiers of data interoperability for sustainable development" (November 2017). Available at https://www.publishwhatyoufund.org/wp-content/uploads/2017/11/JUDS_Report_Web_061117.pdf.
[25] Luis González Morales and Tom Orrell, "Data interoperability: a guide to joining up data in the development sector" (October 2018). Available at https://www.data4sdgs.org/sites/default/files/services_files/Interoperability%20-%20A%20practitioner%E2%80%99s%20guide%20to%20joining-up%20data%20in%20the%20development%20sector.pdf.
[26] See https://medium.com/berkman-klein-center/data-commons-version-1-0-a-framework-to-build-toward-ai-for-good-73414d7e72be.
[27] See https://sdmx.org/.

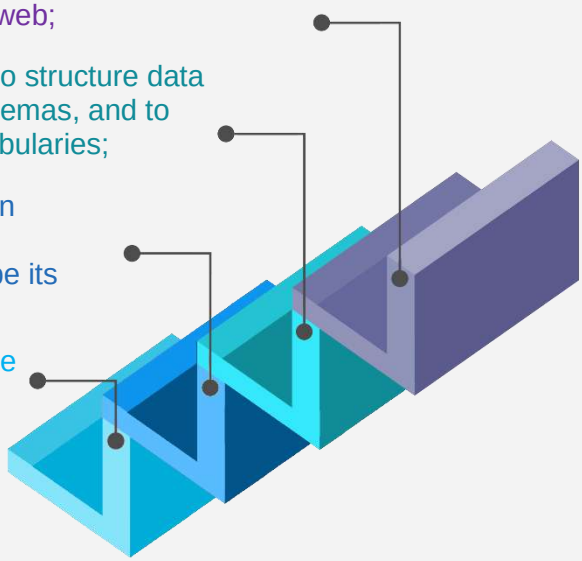
Figure 6. Layers of Interoperability based on the Data Commons Framework

Technology layer: This represents the most basic level of data interoperability, and is exemplified by the requirement that data be published, and made accessible through standardized interfaces on the web;

Data and format layers: These capture the need to structure data and metadata according to agreed models and schemas, and to codify data using standard classifications and vocabularies;

Human layer: This refers to the need for a common understanding among users and producers of data regarding the meaning of the terms used to describe its contents and its proper use.

Institutional and organizational layers: These are about the effective allocation of responsibility (and accountability) for data collection, processing, analysis and dissemination within and across organizations. They cover aspects, such as data-sharing agreements, licences and memoranda of understanding.



Source: Elena Goldstein, Urs Gasser and Ryan Budish, “Data Commons Version 1.0: a framework to build AI for good. Berkman Klein Center, 22 June 2018”.

4.1. CONCEPTS IN THE DATA

39. Most data sources to be integrated are external to the national statistical office. Data from administrative or other non-traditional sources are primarily collected for non-statistical purposes. Consequently, there are often differences in what is required in statistics, such as differences in concepts, coverage, reference period, units and definition of variables.

40. These differences affect the usability of the external data source in the production of statistics specifically with regards to the coverage of the population, the validity of the target concepts, the availability and accuracy of descriptive metadata, sampling error, bias, legal basis for data, data collection methodology, questionnaire design, response burden and confidentiality of the resulting output. These differences need to be clearly explained, well-documented and saved to ensure reuse and improvement of assessments.

41. National statistical offices, through their expertise in data management, can minimize the impacts of these differences and issues with some interventions, such as the development of concordances, pre-processing of data and development of standardized and/or automatic coding approaches. Detailed descriptive metadata are especially needed to assist in the alignment with statistical purposes and the assessment of the quality of the data sources. The following dimensions of quality need to be assessed: accuracy; relevance; consistency; accessibility; comparability; completeness; and timeliness. When other organization's data are used, a statistical organization cannot control all of the decisions on measurements and populations carried out by an external data source provider. A statistical organization needs to understand the design decisions, so that it can determine what to do to turn external data into the statistical information required.

42. Collaboration with data providers is one way to lower the risks. National statistical offices and data providers should agree on using standardized definitions and common classifications where possible. Data providers and statistical organizations have an interest in quality, but the relevant quality aspects and priorities are sometimes different for the production of non-statistical and statistical data. Close collaboration can raise the option to review and revise procedures related to data production in support of data integration.

43. Administrative records are collected to implement various non-statistical programmes concerning legal requirements, such as taxation, housing, pensions, social benefits and trade in goods. Collaboration of the statistical organizations in the preparation of legal documents on establishing and maintaining an administrative source can be very helpful. The approval of the statistical organizations in passing legislation on administrative records may be stated in a statistical act.

44. Control of the methods by which the administrative data are collected and processed often rests with administrative agencies. These agencies specialize in formulating transparent rules and procedures. Statistical organizations have experience in data collection, classifications and data validation. In some cases, the same data are used by several institutions, so continuous collaboration in institutional methodological groups is recommended to develop a system that is satisfactory for both administrative and statistical purposes.

45. An integrated data project or shared platform wherein multiple researchers can access and work across many different data sources will likely present challenges in

understanding datasets conceptually and technically processing the data for the intended purposes. While the concept of a "**learning curve**" is generally considered ubiquitous, resources can be prepared to expedite capacity-building. In addition to the development of a **metadata and resources library**, the national statistical office or an integrated data project could host digital spaces to promote the sharing of knowledge and experiences, and more importantly "questions and answers" exchanges between researchers and/or data suppliers. The latter is of special interest to researchers and data users, as it addresses actual concerns and information gaps.

4.2. IDENTIFIERS

46. An important requirement for data integration is connectivity. This can be done most easily under a unified identification system across different sources. In many countries, national unified identity systems exist for persons, businesses, farmers and addresses (or geo codes). The identity numbers are often anonymized and translated into statistical identity numbers for privacy protection in the statistical production.

47. In cases in which a unique identifier is not available, lacks sufficient quality or coverage to be relied on, or not used by all data providers, it is more difficult to link different sources. If the sources contain unique identifiers, the integration is directly achieved via these identifiers; otherwise, it is necessary to define and prepare a procedure for pooling records using a set of key variables that are common among data sources. Further information on this is available in chapter 6.

48. A number of approaches can be applied to protect the privacy of unique identifiers. One of them is the use of generated temporary unique identifiers produced on user request based on the unique IDs, such as temporary identifiers in mobile transactions. Another technique is to use a one-way hashing algorithm. This approach, if properly implemented, can be used to create new identifiers that are difficult to "reverse" to discover the original identifier (see Introduction to The Hash Function as A Personal Data Pseudonymisation Technique).[28] An approach to replacing the identifiers is also described in the Statistics New Zealand *Data Integration Manual: 2nd edition*.[29]

49. Results of the ESCAP 2020 survey (DI-CAS) show that many of responding countries have a unified personal/business identity system in place. Less than half of responding countries hold a unified address system. Although identifiers are established for most countries that responded to the survey, for around half of responding organisations, data custodians in their country had restrictions on providing identifiers to the national statistical office.

[28] Agencia Expañola Protección datos and the European Data Protection Supervisor, "Introduction to the hash function as a personal data pseudonymistion technique" (October 2019). Available at https://edps.europa.eu/sites/default/files/publication/19-10-30_aepd-edps_paper_hash_final_en.pdf.

[29] Statistics New Zealand, *Data Integration Manual: 2nd edition* (Wellington, New Zealand, Statistics New Zealand, 2013). Available at https://web.archive.org/web/20200212144854/http://archive.stats.govt.nz/methods/data-integration/data-integration-manual-2edn.aspx.

4.3. PRIVACY AND CONFIDENTIALITY

50. Privacy refers to freedom from intrusion into one's personal information. Confidentiality concerns about personal information shared with others that generally cannot be divulged to others without the express consent of the individuals. Confidentiality means that the information can be accessed only by authorized individuals.

51. *Social licence to operate* is a concept stemming from the corporate social responsibility framework and knowledge. It represents the ongoing and sustained public acceptance and trust of any activity, project, or organization, beyond legal compliance. Because of growing awareness of pernicious impacts from government or private endeavours, social concerns and resistance may arise and put into question the intended operation. In the context of a data integration project or platform, it is closely linked to public concerns over individuals' privacy and confidentiality and enhanced surveillance techniques. These concerns, if disregarded, may lead to conflicts and social resistance, for example, populations refusing to participate in surveys and litigation over access and control of personal data.

52. Some key points to note from the established body of knowledge on *social licence to operate* concern the basic principles to manage it: (a) identifying stakeholders by understanding potential impacts and perceptions; (b) identifying key areas of concern, namely human rights, privacy and surveillance; and (c) developing action and communication plans to address and mitigate impacts. Further information about the concept of a social licence to operate can be found in a presentation given by Ian Thomson and Susan Joyce at the PDAC Convention in 2008.[30] A relevant standard (ISO 26000: social responsibility) is also available.[31]

53. Protecting personally identifiable information (PII) is essential for personal privacy, data privacy, data protection, information privacy and information security. PII typically refers to information that can be used to distinguish or trace an individual's identity, either by itself or in combination with other personal or identifying information that is linked or linkable to an individual. PII may include name, address, email, telephone number, date of birth, passport number, fingerprint, driver's licence number, credit or debit card number and social security number, among others. (personally identifiable information (PII) by Corinne Bernstein).[32] PII is a well-known concept in international standards related to databases containing personal information that must be protected. People involved in data integration need to have a comprehensive understanding of this concept and what constitutes it.

54. Integrating, holding and storing more data sources can increase disclosure risks that need to be managed carefully. The ability to integrate data sources depends on the trust of observation units: persons; households; enterprises; agricultural holdings; and other entities. This means that respondents, administrative agencies and other data providers should only share their data if they are convinced that the confidentiality of the data and identity is ensured, PII is protected and the shared data are only used for statistical purposes. To assure public acceptance, privacy and confidentiality rules must be clear.

---

[30] Ian Thomson and Susan Joyce, "The social licence to operate: What it is and why does it seem so different", presentation at PDAC Convention, Toronto, Canada, March 2008. Available at https://oncommonground.ca/wp-content/downloads/PDAC_2008_Social_Licence.pdf.

[31] See https://www.iso.org/iso-26000-social-responsibility.html for a relevant standard.

[32] Corinne Bernstein, "Personally identifiable information (PII)", Techtarget Network (February 2020). Available at https://searchsecurity.techtarget.com/definition/personally-identifiable-information-PII.

55. Many countries have enacted a personal data protection act, which sets the rules on processing personal data in a way that the legal rights of individuals concerning privacy and the integrity of individual's data are not violated.

56. The protection (safeguarding) of confidentiality aims to ensure that disseminated data do not allow direct identification (via identifiers) or indirect identification (by any other means). To this end, appropriate statistical disclosure methods are needed to ensure PII protection in compliance with the legislation. A data-sharing standard operating practice document should be prepared to avoid breach of confidentiality by people involved in integrating the data.

57. Most national statistical offices are prohibited by law from breaching confidentiality and already have in place mechanisms to protect privacy and confidentiality. They also hold detailed information about individuals collected through censuses and surveys. National statistical offices can, therefore, be considered as a trusted party, and data flow of individual records needs to be one-way, towards the national statistical office. Data flowing to other organizations may need to be protected by perturbation, aggregation or other procedures.

58. All data integration projects must take into account the details of the datasets and suppliers. No new data integration should happen without a comprehensive analysis of its potential risks and benefits. The preparation and publication of privacy and confidentiality impacts assessments for the integration of datasets is a best practice approach to detect, evaluate and mitigate potential risks. The aim of these documents is to inform the public of the intention to integrate one or more datasets from which personal information will be handled and to clearly demonstrate that the benefits of doing so outweigh the risks. These documents should be drafted in response to and be guided by the principles of the national privacy and confidentiality authority (if such exists) or provide a per case privacy and confidentiality agreement, or refer to international standards, such as ISO/IEC 29151[33] (information technology — security techniques — code of practice for personally identifiable information protection), and ISO/IEC TR 38505-2[34] (information technology — governance of IT — governance of data — Part 2: implications of ISO/IEC 38505-1 for data management).

59. Below are suggested sections for a privacy and confidentiality assessment document:

- An introduction to the document, clearly outlining the scope and coverage of the intended operation, namely integration of one or more datasets;
- A general description of the datasets and their contents;
- A summary of the expected benefits of carrying out the data integration;
- A detailed analysis and assessment of risks to privacy and confidentiality, their rationale and a thorough description of safeguards and mitigation procedures to protect PII;
- The decision made and all considerations taken into account as a result of conducting the privacy and confidentiality assessment.

[33] See https://www.iso.org/standard/62726.html.
[34] See https://www.iso.org/standard/70911.html.

60. Examples from Statistics New Zealand are available on its webpage in an article entitled “Data integration projects and privacy impact assessments”.[35]

61. If a statistical system is centralized and the national statistical office plays a leading role in data integration efforts, concerns about public acceptance of data integration is likely to be less as compared to decentralized statistical systems. On the other hand, in a centralized statistical system, data partners, such as registrars may be less familiar with data considerations for statistical use. Accordingly, different statistical system types may require different regulations and considerations for the management of data integration.

[35] See https://web.archive.org/web/20190122153642/http:/archive.stats.govt.nz/methods/data-integration/data-integration-projects.aspx.

Chapter 5

QUALITY

62. Quality assessment is an important issue in statistical production. Many organizations and international and national initiatives focus on various aspects of quality, and some of them explicitly consider processes related to data integration. The following dimensions of quality[36] need to be assessed: relevance; accuracy; timeliness; accessibility; coherence; and interpretability. ESSnet Komuso in the final version of Quality Guidelines for Multisource Statistics (QGMSS),[37] categorizes the statistical sources of errors arising in the multisource processes from the survey component, the administrative data component or both (figure 7), and shows how each type of errors and other factors are linked to dimensions of quality (figure 8).

63. A useful reference for the measures of quality in data integration steps is a section on quality indicators for the GSBPM.[38] In this work, indicators to evaluate the quality of standard linkage procedures are proposed.

64. When integrating survey data and administrative data, the ESSnet Komuso in Quality of multisource statistics[39] provides useful documents to do the following:

    - Take stock of the existing knowledge on quality assessment and reporting and review it critically in order to produce recommendations of the most suitable approaches;
    - Develop new indicators for the quality of the output based on multiple sources;
    - Produce a methodological framework for reporting on the quality of output;
    - Produce indicators relating to the quality of frames themselves and the data whose production is supported by frames;
    - Produce recommendations for updating the *ESS Standard and the ESS Handbook for Quality Reports*.

65. The ESSnet Komuso Quality Guidelines for Multisource Statistics (QGMSS) is a practical and applicable guide supporting the design and implementation of multisource statistics within a comprehensive quality framework; a complete overview of the set of output quality measures, supported by applications and computation details, are provided in the report entitled “Complete overview of quality measures and calculation methods (QMCMs)”.[40]

66. The outputs of the Methodologies for an Integrated Use of Administrative Data in the Statistical Process (MIAD) project (deliverables series B) provide a generic framework to assess the quality of the administrative data at the input stage, quality indicators for the discovery phase and acquisition phase, and a guide to reporting the usability of an administrative data source.

---

[36] See https://ec.europa.eu/eurostat/cros/system/files/Guide%20to%20report%20the%20usability%20of%20an%20ADS.pdf.

[37] ESSnet KOMUSO, Work Package 1, Quality Guidelines for Multisource Statistics (QGMSS), Version 1.1. Specific Grant Agreement No. 3 (SGA-3) (October 2014). Available at https://ec.europa.eu/eurostat/cros/system/files/qgmss-v1.1_1.pdf.

[38] See https://statswiki.unece.org/display/GSBPM/Quality+Indicators.

[39] European Commission, ESSnet on quality of multisource statistics – Komuso. Available at https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en.

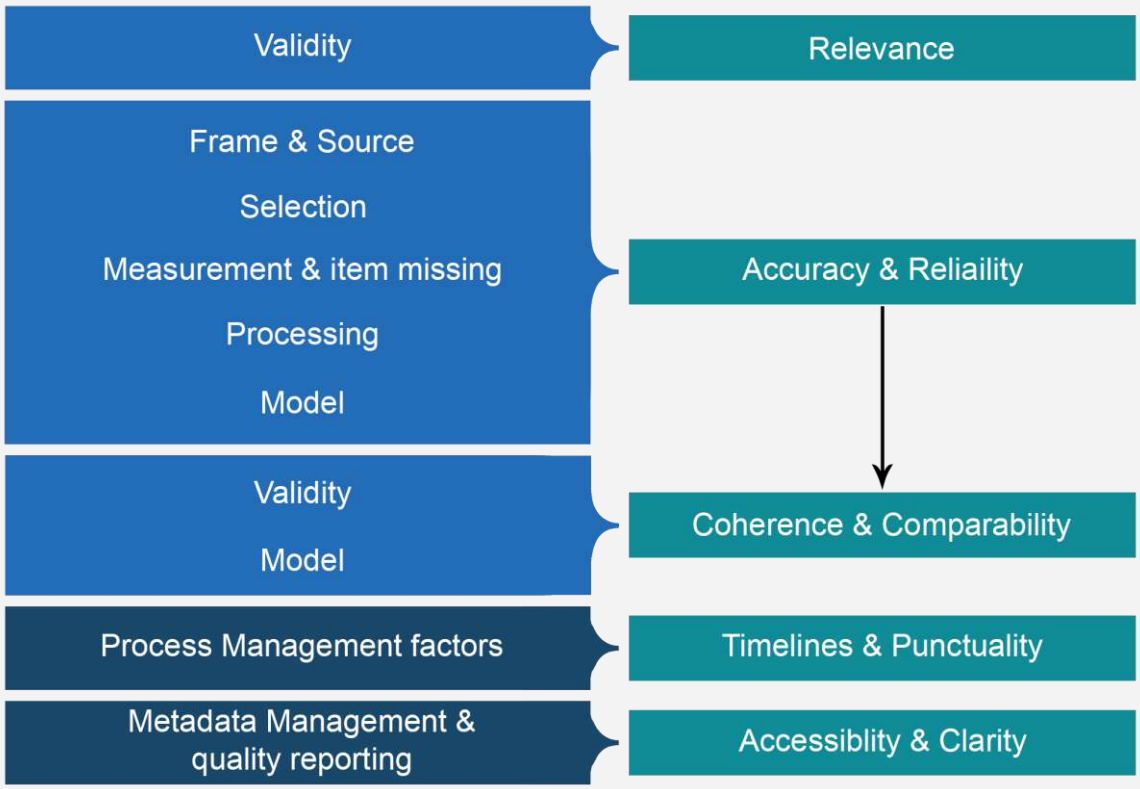
[40] Eurostat, “Complete overview of quality measures and calculation methods (QMCMs)” (October 2019). Available at https://ec.europa.eu/eurostat/cros/system/files/qmcms_examples_overview_1.pdf.

Figure 7. Main sources of errors in multisource statistics

Error category	Type of error included	Survey	Administrative sources
Validity error	Specification error	X	
	Relevance error		X
Frame and source error	Under-coverage	X	X
	Over-coverage	X	X
	Duplications	X	X
	Misclassification in the contact variables	X	
	Misclassification in the auxiliary variables	X	X
Selectivity error	Sampling error	X	
	Unit non-response	X	
	Missing units in the accessed data set		X
Measurement error and Item missingness	Arising from: respondent, questionnaire, interviewer, data collection	X	
	Fallacious or missing information in administrative source		X
Processing error	Data entry error	X	
	Coding or mapping error or misclassification	X	X
	Editing and imputation error	X	X
	Identification error		X
	Unit error		X
	Linkage errors	X	X
Model error	Editing and imputation error, record linkage error, etc.	X	X
	Model based estimation error (small area estimation, seasonal adjustment, structural equation modelling, Bayesian approaches, capture-recapture or dual system estimation, statistical matching, etc.)	X	X

Source: ESSnet KOMUSO, Work Package 1, Quality Guidelines for Multisource Statistics (QGMSS), Version 1.1. Specific Grant Agreement No. 3 (SGA-3) (October 2014).

Figure 8. Main type of errors and other factors influencing output quality dimensions



Source: ESSnet KOMUSO, Work Package 1, Quality Guidelines for Multisource Statistics (QGMSS), Version 1.1. Specific Grant Agreement No. 3 (SGA-3) (October 2014).

67. At the national level, statistical organizations recognize the necessity of developing a framework for assessing quality in the usage and integration of different data sources. The following resources are recommended for this purpose:

- *Statistics Canada Quality Guidelines* (2019)[41]
- Checklist for the quality evaluation of administrative data sources of statistics Netherlands (2009)[42]
- Guidelines for the quality of statistical processes that use administrative data of Istat (2016)[43]

[41] See https://www150.statcan.gc.ca/n1/en/catalogue/12-539-X.
[42] Piet Daas and others, "Checklist for the quality evaluation of administrative data sources of statistics Netherlands", discussion paper (The Hague/Heeren, Netherlands, Statistics Netherlands, 2009). Available at http://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60.
[43] ISAT, "Guidelines for the quality of statistical processes that use administrative data" (August 2006). Available at https://www.istat.it/en/files/2013/04/Linee-Guida-v1.1-Versione-inglese.pdf.

- Guide to reporting on administrative data quality (Statistics New Zealand (2020))[44]
- Quality Assurance of Administrative Data (United Kingdom Statistics Authority (2015))[45]
- Data Quality Framework (Bank of England (2014))[46] and ECB Statistics Quality Framework (European Central Bank (2008))[47]

68. The quality assessment framework of Statistics New Zealand, including the quality indicators, is described in the Guide to reporting on administrative data quality. This quality framework is based on a two-phase life-cycle method model for integrated statistical microdata developed by Li-Chun Zhang[48] (figure 9), which expands the total survey error paradigm to include administrative data.

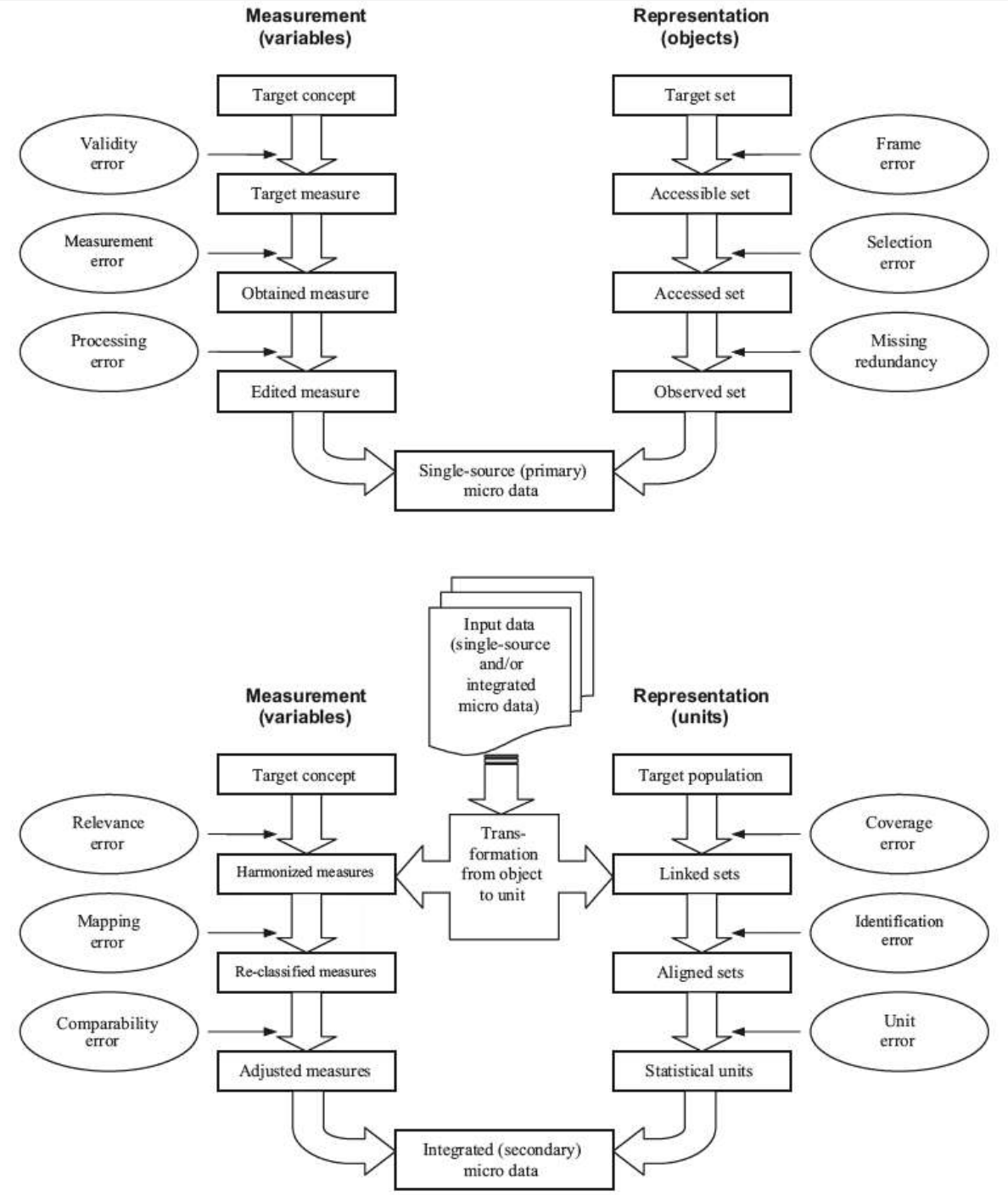
69. The framework enables users to gain an understanding of sources of errors in both single-source and integrated micro data. The two-phase life-cycle model assists in determining the associated methodological and operational issues that may have an impact on quality resulting from producing statistical information from linked administrative data sources.

70. In phase 1, the quality of an input data source intended to be used in the production of a statistical product is assessed. A statistical organization needs to understand the design decisions carried out by the producers of the source to determine methods to turn the data into the required statistical information. Quality of the input data source is assessed against the purpose for which it was collected. For a survey dataset, the purpose is defined for a statistical target concept and target population. For an external data source, the entries or "objects" in the dataset may be people or businesses, but they can also be transaction records, or other events of relevance to the collecting agency. At this stage, evaluation is entirely based on the dataset itself, and does not depend on what a statistical organization intends to do with the data. Quality issues in the input data source will flow through into any use of the data in the production of a statistical product.

71. In phase 2, the difficulties arising from taking variables and objects from source datasets and using them to measure the statistical target concept and population of interest to a statistical organization are categorized. During this phase, the statistical organization considers what they want to do with the data and determines how well the source datasets match with what they would ideally be measuring.

[44] Statistics New Zealand, "Guide to reporting on administrative data quality", 18 November 2020. Available at https://www.stats.govt.nz/methods/guide-to-reporting-on-administrative-data-quality.
[45] United Kingdom Statistics, "Quality assurance of administrative data – Setting the standard", version 1 (January 2015). Available at https://osr.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-settingthestandar_tcm97-44370.pdf.
[46] Bank of England, Statistics and Regulatory Data Division, *Data Quality Framework* (London, Bangkok of England, March 2014). Available at https://www.bankofengland.co.uk/-/media/boe/files/statistics/data-quality-framework.
[47] European Central Bank, "ECB Statistics Quality Framework (SQF)" (April 2008). Available at https://www.ecb.europa.eu/pub/pdf/other/ecbstatisticsqualityframework200804en.pdf?a7dfa6c0d9310632f050e1e533fe9586.
[48] Li-Chung Zhang, "Topics of statistical theory for register-based statistic and data integration", *Statistica Neerlandia*, vol. 66, No1 (February 2012) pp. 41−63.

Figure 9. Two-phase life-cycle method model for integrated statistical microdata developed



Source: Li-Chung Zhang, "Topics of statistical theory for register-based statistic and data integration", Statistica Neerlandia, vol. 66, No1 (February 2012) pp. 41−63.

72. The quality assessment involves three steps, as below.

Step 1: Initial metadata collation:

Basic information is collected about each of the source datasets used in the validation project. The information is related to the source agency, purpose of the data collection, populations, variables and timeliness of the data.

Step 2: Phase 1 evaluation:

Errors occurring in phase 1 of the quality framework are determined and categorized for each source dataset. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the phase 1 flow chart, as shown in figure 9.

Step 3: Phase 2 evaluation:

As for the previous step, errors arising in phase 2 of the quality framework are listed and examined in a similar way, taking into account the dataset(s) being integrated to produce the final output. These errors are considered with respect to the intended statistical target concepts and population. The effects of phase 1 errors on the creation of statistical units, or the particular details of the misalignment between concepts on different datasets, must be understood.

73. The Statistics New Zealand Guide to reporting on administrative data quality provides a metadata information template that encourages thinking about the key aspects of quality in an organized manner. It is also a convenient way to record a standard set of information to compare different datasets. The basic information required are: name of data source agency; purpose of data collection; time period covered by the data; the population (target and actual) of the dataset; the reporting units; a short description of key variables; and the timing and delay information and method of collection.

74. For the integration of big data sources in the statistical production, several global initiatives have indicated the potential of these new data sources and the quality issues related to several aspects. First, the change of paradigm imposed by the new sources, compared to the traditional sample surveys, moves attention from the well-studied sampling errors to the non-sampling errors, so the population coverage and the self-selectivity of the observations become the most recognized and investigated issues. In 2014, the ECE Big Data Quality Task Team issued “A suggested framework for the quality of big data”.[49]

[49] ECE Big Data Quality Task Team, “A suggested framework for the quality of big data” (December 2014). Available at https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2#:~:text=The%20Big%20Data%20Quality%20Framework,framework%20developed%20by%20Statistics%20Netherlands.

Chapter 6

METHODS AND TOOLS

75. Data integration procedures differ in terms of the types of data sources to be combined, the characteristics of datasets, such as the coverage and overlapping of datasets through the data sources, the micro or macro levels of data, the existence and usability of unique identifiers, and the purposes of combining data.

76. Data integration methods can be divided into two main groups, record linkage methods and statistical matching methods, which, in turn, can be further divided into several subgroups and categories from different perspectives.

77. The literature available on these methods is increasing, covering their subprocesses, advantages and disadvantages, mathematical bases, and tools. Insights on Data Integration Methodologies[50] and chapters 5, 6 and 7 of the Statistics New Zealand *Data Integration Manual 2nd Edition* give useful and detailed information about the methodological aspects of data integration.

78. A brief overview of data integration methods along with some of their most notable features and some of the tools used in official statistics to carry out data integration are presented here.

6.1. RECORD LINKAGE

79. *Record linkage* refers to the identification and combination of records corresponding to the same entities – for example, persons, enterprises, dwellings and households – throughout two or more data sources. Record linkage methods can be classified into two branches:

    - *Deterministic matching* (or exact matching) is when a formal decision rule – usually the coincidence (or mismatch) of the unique identifiers that correspond to the same units in two or more data sources – is applied to determine whether a pair of records is a match.
    - *Probabilistic matching* is when strict decision rules are not applicable. Instead, complex probabilistic decision rules are established based on a set of key variables that are common in the datasets to be integrated to be able to accept or refuse matches on a probabilistic basis.

80. In recent years, a variety of machine-learning techniques have been used in record linkage. *Machine*-learning methods are categorized into two main groups: supervised learning, when a training set is available; and unsupervised learning. Using machine-learning techniques has been shown improved the accuracy of record linkage and made it possible to increase the number of linked records.

81. Because of the similarities between deterministic and probabilistic matching methods, namely that both are based on the matching of key variables, they share common features. Both the deterministic and probabilistic matching procedures, can lead to linkage errors when false matches (or false positives) are interpreted as real ones or false unmatches

[50] Eurostat, *Insights on Data Integration Methodologies* (Luxembourg, European Communities, 2009). Available at https://www.istat.it/it/files/2015/04/Insights-on-Data-Integration-Methodologies.pdf.

(or false negatives) that is, real matches are not recognized as such. Moreover, both of them consist of similar phases, as explained below:

- ***Pre-processing:*** This phase includes harmonizing datasets in terms of, definitions, coverage, reference period, classification and coding and data formats, among others. A brief methodology on how to align data sources with statistical requirements should be available. During this phase, erroneous or suspicious data are detected and dealt with, format is standardized for the fields stored in different formats and coding is made consistent across files and variables. In addition, common variables are identified, and key variables for matching are chosen. Key variables should all be in standard format.
- ***Linkage:*** Regardless of the linking method being used, this phase will result in one of the following:
    - Match (same entity);
    - Unmatch (different entities);
    - Uncertain match (unable to decide – possible match).
- ***Post-linkage (manual review of unlinked records):*** After linkage, unlinked records and uncertain matched data are reviewed manually. It may be required to use other datasets to deal with unmatched or uncertain matched data.
- ***Data analysis:*** During this phase, quality of matching is evaluated and quality measures, such as rate of linkage errors, are estimated.

Data Integration in the compilation of cohort-based marriage and divorce indicators in Singapore[51] illustrates practical aspects of above-mentioned phases.

82. *Deterministic matching* is considered as the ideal case of record linkage due to the existence of a unique identifier, such as the social security number of persons, fiscal code of enterprises and geocodes of addresses, which usually assures an error-free, one-to-one matching of records with the same identifier that belongs to the same entity. For this reason, there is considerably less literature on this method as compared with other methods. However, some challenges can emerge during the application of this method. A possible difficulty is that unique identifiers can also be affected by errors occurring, for instance, during the data collection or data capture process. In addition, there may be missing values in some of the data sources. Identifying records in base registers – in the "spines of integration" – may be a useful solution in order to obtain or check unique identifiers.

83. *Probabilistic matching* is a more complex approach. Instead of unique identifiers, softer key variables are used, such as the name, date of birth, address or other variables describing the units of the target population. These variables are more likely to be affected by data collection or data capture errors, or they are often recorded in different formats, making the comparison of them more complicated. In such cases, the pre-processing phase plays a crucial role that can strongly affect the results of a record linkage exercise.

84. Language characteristics can introduce different issues to record linkage; for example, RTL (right to left) languages may differ from LTR (left to right) languages. To link the records based on text fields, all potential problems raising from type of writing should be considered in the pre-processing phase. Solutions may include such steps as normalizing,

[51] Statistics Singapore, "Data Integration in the compilation of Cohort-Based Marriage and Divorce Indicators in Singapore", presentation. Available at https://www.unescap.org/sites/default/files/Session3_Singapore_DI-CoP_WS_24-27Nov2020.pdf.

removing punctuations and extra space characters, spelling correction, removing stop words, stemming and tokenizing. Another step might be to create a document-term matrix. In this step, one method is to create unigram or n-gram variables in order to count the frequency or presence or absence of the words in the given text.

85. The complex mathematical basis of probabilistic record linkage and probabilistic decision rules go back to the ground-breaking works of H. B. Newcombe and others[52] and I. P. Fellegi and A. B. Sunter,[53] who formalized the theory of probabilistic matching based on the assumption of conditional independence. Even today, this method serves as the basis of record linkage applications. Other probabilistic record linkage techniques are that of M. A. Jaro in 1989,[54] which was further developed by W. E. Winkler,[55] or the distance-based record linkage method as described by D. Pagliuca and Giovanni Seri.[56]

[52] H. B. Newcombe and others, “Automatic linkage of vital records”, Science, vol.130, No 3381 (1959) pp. 954–959.

[53] I. P. Fellegi and A. B. Sunter, “A theory for record linkage”, *Journal of The American Association Statistical Association*, vol. 64, No. 328 (1969) pp. 1183–1210.

[54] M. A. Jaro, “Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa”, Florida, *Journal of the American Statistical Association*, vol. 84, No. 406, pp. 414–420.

[55] W. E. Winkler (1995), “Matching and record linkage. Business Survey Methods”, (New York, J. Wiley and Son).

[56] D. Pagliuca and Giovanni. Seri, “Some results of individual ranking method on the system of enterprise accounts annual survey”, Espirit SDB Project, Deliverable MI-3?D2.

6.2. STATISTICAL MATCHING

86. *Statistical* matching (or synthetic matching) involves the integration of data sources with usually distinct samples from the same target population in order to study and extend information on the relationship of variables not jointly observed in the datasets. According to Aura Leulescu and Mihaela Agafitei in a paper written in 2013, the main difference from record linkage is that "record linkage" deals with identical units, while statistical matching deals with "similar" units.[57] In practice, matching procedures can be regarded as an imputation problem of the target variables from a donor to a recipient survey. The statistical matching situation is usually described with a recipient data source A containing X and Y variables and donor data source B with X and Z variables. That is the statistical matching itself is imputing Z variable in data source A using the common variable X. *Statistical Matching: Theory and Practice*[58] is one of the most important resources available on statistical matching.

Figure 10. Statistical matching illustration

	Y	X	Z
Data source *A*		Common variable	missing
Data source *B*	missing		

Source: Eurostat, (2014). "Micro-Fusion", In Methodology of Modern Business Statistics, module (26 March 2014).[59]

87. Statistical matching methods are categorized in the specialized literature from different angles:

- *The micro approach* is aimed at constructing a complete (containing all variables of interest) and synthetic (comprise of not directly observed units) micro-level dataset.
- *The macro approach* is aimed at integrating data sources to facilitate the estimation of the parameters of interest as the correlation or regression coefficients and the contingency tables of not jointly observed variables at the macro level.

88. From another angle, the *micro* and *macro* approaches can be parametric or non-

[57] Aura Leulescu and Mihaela Agafitei, "Statistical matching: a model based approach for data integration", Eurostat Methodological and Working Papers (2013). Available at https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF.

[58] Marcello D'Orazio, Marco Di Zio and Mauro Scanu, *Statistical Matching: Theory and Practice* (Hoboken, New Jersey, John Wiley and Sons, 2006).

[59] Available at https://ec.europa.eu/eurostat/cros/system/files/Micro-Fusion-08-T-Statistical%20Matching%20v1.0_2.pdf.

parametric and a mix of them can also be applied for the *micro* approach:

- The *parametric approach* is usually based on the normality assumption of data. In this case, a specified model is needed for the joint distribution of the variables that can lead to misspecification (usually maximum likelihood).
- The *non-parametric approach* is applied when data do not hold the normality assumption. This approach is more flexible than the parametric one when variables are of different types (usually hot-deck techniques).
- A *mix* of the parametric and the non-parametric approaches can be applied for carrying out micro-level matching: "first a parametric model is assumed and its parameters are estimated, then a synthetic dataset is derived through a non-parametric micro approach. In this manner the advantages of both parametric and non-parametric approaches are maintained: the model is parsimonious while non-parametric techniques offer protection against model misspecification".[60]

89. Furthermore, approaches can be distinguished in accordance with the availability of information on variables that are not observed jointly:

- Approaches that assume the *conditional independence* of variables (originally all the micro-, macro-, parametric and non-parametric, and mixed methods were based on the conditional independence assumption). Conditional independence of the target variables given the common variables means Y and Z are independent once conditioning on X variables.
- Approaches in which *auxiliary information* is available from one of the datasets or from a third data set in which variables are jointly observed.
- In the case of *uncertainty*, no assumptions are made, and no joint information is available on the variables, so uncertainty analysis techniques are applied usually at the macro level.

90. Micro approach statistical matching is an imputation technique. This means that statistical matching is estimating missing Z variables. If a statistically matched dataset is used as if the X, Y and Z variables were jointly observed, it ignores the fact that the Z variables are just estimates. Ignoring this uncertainty will lead to incorrect variance estimates. According to D. B. Rubin multiple imputation is a way to obtain correct variance estimators. The core idea of multiple imputation is to not only impute the data once but to draw *m* times (often about five) from a set of plausible values.[61] The *m* imputed datasets are then separately analysed the same way complete data would be analysed, and the results are combined using the rules of Rubin. The rules take both the within and the between variance of the imputed values into account, which leads to more appropriate variance estimates. In statistical matching, a basic approach to obtain these multiple imputed datasets (that is close to the idea of Rubin) is to find the *k* nearest donor neighbours for each recipient and choose at random and with replacement *m* of these *k* donors. *Multiple Imputation and its Application*[62] is a good source on this.

---

[60] Marcello D'Orazio, "Statistical matching and imputation of survey data with StatMatch" (December 2017). Available at https://cran.r-project.org/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf.

[61] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys* (New York, John Wiley & Sons, Inc., 1987).

[62] James R. Carpenter and Michael G. Kenward, *Multiple Imputation and its Application* (Hoboken, New Jersey, John Wiley & Sons, Inc., 2012).

6.3. SOFTWARE TOOLS

91. The ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data in the "Report of WP3. Software tools for integration methodologies"[63] reviews some existing software tools for the application of probabilistic record linkage and statistical matching methods. The report lists the following software tools and compares them with each other:

    - *Record linkage tools:* AutoMatch, Febrl, Generalized Record Linkage System (GRLS), LinkageWiz, RELAIS, DataFlux, Link King, Trillium Software, Link Plus, Record Linkage (R codes)
    - *Statistical matching tools:* StatMatch (R code), SAM WIN, SAS code, SPlus code

92. GitHub also briefly overviews data matching software tools,[64] which are open source and/or freely available. Figure 11 gives a dense overview of data matching software properties. The properties evaluated are Application Programming Interface (API), Graphical user interface (GUI), linking, deduplication, supervised learning, unsupervised learning, and active learning.

93. In the 2020 ESCAP survey (DI-CAS), organizations were asked about the tools (applications and software) they use for linking and/or matching data. The most widely used tools (applications, software) for linking and/or matching were Excel, SQL, SPSS, and R.

[63] Available at https://ec.europa.eu/eurostat/cros/content/deliverables-wp3software-di-isad-wp3_en.
[64] See https://github.com/J535D165/data-matching-software#jedai.

Figure 11. Dense overview of data matching software properties

Software	API	GUI	Link	Dedup	Supervised Learning	Unsupervised Learning	Active Learning
AtyImo	PySpark	✗	✓	✓	✗	✗	✗
Dedupe	Python	✗	✓	✓	✓	✗	✓
fastLink	R	✗	✓	?	✗	✓	✗
FEBRL	Python	✓	✓	✓	✗	✗	✗
FRIL	Java	✓	✓	✗	?	✓	✗
FuzzyMatcher	Python	✗	✓	✗	✗	✓	✗
JedAI	Java	✓	✓	?	✓	?	?
PRIL	SQL	✗	✓	?	?	?	?
Python Record Linkage Toolkit	Python	✗	✓	✓	✓	✓	✗
RecordLinkage (R)	R	✗	✓	✓	✓	✓	✗
RELAIS	✗	✓	✓	?	?	✓	✗
ReMaDDer	✗	✓	✓	✓	✗	✓	✗
Splink	PySpark	✗	✓	✓	✗	✓	✗
The Link King	✗	✓	✓	✓	?	✓	✗

✓ Yes/Implemented ✗ No/Not implemented ? Unknown

6.4. OTHER METHODOLOGICAL CONSIDERATIONS

94. A common issue with linked datasets is inconsistencies in the records linked. In cases in which inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining the point from which the inconsistency arose. It is expected that as the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records grows. This may also add to the amount of editing required for the linked datasets.

95. Issues to be addressed by an editing strategy for linked datasets can be summarized by the ability to (a) edit inconsistencies from the same unit across different sources, (b) treat erroneous and missing variables in a record and (c) ensure consistency in variables across a record for a time period and over time.

96. Sources of potential bias have been identified with regard to integrating datasets. These include the following:

    - Coverage and conceptual issues may apply for some groups of a population, so care should be taken in generalizing results.
    - Some variables have the potential to affect the quality of linking and may be a source for potential bias in carrying out analysis on resulting datasets. Investigations on linkage rates across different subpopulations may be required.
    - Using linked datasets (even for validation purposes) may result in a break in the data series that needs to be managed.

97. Extreme care should be taken in backwards and forward casting of linked data, especially for longitudinal data. Because of data quality issues, an individual may link correctly in one quarter but not in another. A weight may be needed to adjust for missed links in linked datasets.

98. Methods to better estimate linkage errors are required to determine models appropriate to account for linkage errors. Linkage errors contribute to potential coverage errors in the resulting target population. Focus should also be placed on efforts to create statistical units from integrated datasets whenever external data sources have been used, because units may be defined differently in the external dataset.

99. Data sourced externally may be adversely affected by measurement errors. These errors propagate when the data are integrated with other data sources to produce a statistical output. Accordingly, target concepts used in a dataset sourced externally should be well understood before being used in the production of official statistics.

Chapter 7

COMMUNICATING INTEGRATED DATA

100. To communicate means to make something known, to impart and to transmit information. Importantly, communication is a two-way process, involving information and messages going back and forth between a sender and receiver.

101. Communication goes beyond “dissemination”, which can be a one-way activity to publish statistics online or in print. It involves talking with stakeholders and data users to capture and understand data needs, finding the most efficient way to meet their needs, explaining the methods and results and seeking their feedback to continually improve statistical products.

102. Communication is essential throughout the entire data cycle. Communication activities include the following:

- Understanding needs and demands for data;
- Establishing data-sharing agreements;
- Explaining methods;
- Presenting results;
- Seeking from and giving feedback to suppliers and users, and so on.

7.1. KEY AUDIENCES

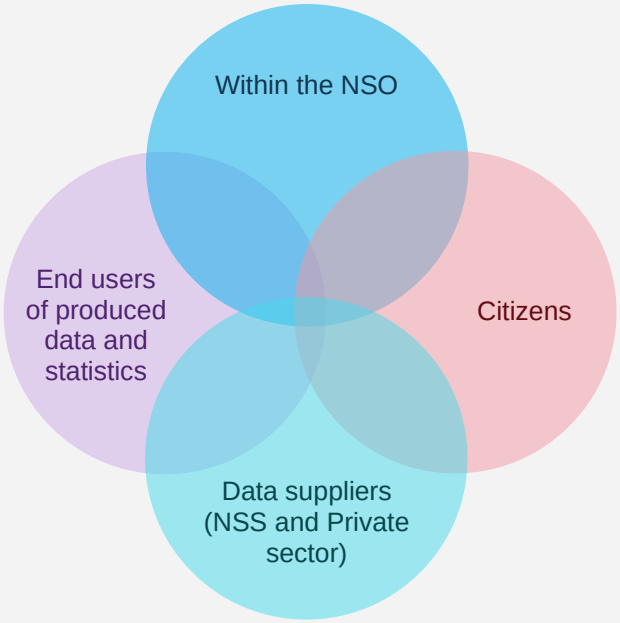
103. Producing data and statistics by integrating data from different sources is a relatively new practice to most of the national statistical systems. This highlights the importance of communication to keep all stakeholders informed and ensure integrated data reach their intended audiences.

104. Regarding communicating integrated data, there are several target audiences to keep in mind (figure 12):

- **Staff within the national statistical office** who need to know national and organizational policies for data integration and what the national statistical office is doing and planning in this space.
- Other **data suppliers across the national statistical system** that the national statistical office is working with to attain access to their data sources and use them in data integration products. This audience needs information about how to standardize their data, what data-sharing agreements are in place, and how they should format and transmit the data. They also need to know how the data are being used by the national statistical office and how confidentiality is being protected.
- **End-users of the integrated data** are a key audience that have specific communication needs. They require information on why and how the data integration has been done, the strengths and limitations of the final product, and what they need to know in order to interpret and use the data effectively. The types of data users vary from high-level decision makers who need clear and succinct information to researchers and academia who are experts in data and statistics and need technical guidance rather than descriptive analysis.

- **Citizens and respondents** are an extremely important audience to consider. They may not use the integrated data products directly but have contributed their information and want to ensure the safe and ethical treatment of private information by the government.

Figure 12. Key audiences



***Note*:** NSO, national statistical office; NSS, national statistical system.

7.2. WHAT IS DIFFERENT ABOUT COMMUNICATING INTEGRATED DATA?

105. Three main aspects make communicating integrated data slightly different from other sources of official statistics.

- Data integration relies on building relationships, which is dependent on good communication. Effective communication is crucial for creating win-win partnerships with policymakers, other governmental agencies and the private sector. Much time needs to be spent explaining the needs for and benefits of data integration and the role that other agencies need to play in regularly providing the required data.
- Data integration uses new, emerging, and sometimes complex methods that need to be explained. The methods used for data integration require good communication skills and practices. Written communication is needed to document the approaches and provide guidelines for others that need to contribute to or to replicate integration methods. National statistic offices and national statistical systems must be able to explain the value proposition of integration so that key audiences understand why it is being done and the benefits that will be realized as a result. It is also necessary to explain (communicate) about data integration methods to non-technical users in layman terms verbally and in writing.
- Mitigating risks to trust and credibility is essential. Citizens need to be informed that data integration is being done and that their concerns about data use and confidentiality are being addressed. National statistical offices should communicate that statistical confidentiality is an overriding principle and always a priority in the production of official statistics. According to a paper about the future role of national statistical offices in Europe, national statistical offices should be leading discussions to clarify the difference between concepts of privacy, security and (statistical) confidentiality.[65]

[65] United Nations, Economic Commission for Europe, “Implementation of the new role of national statistical offices at the time of expanded possibilities”, Conference of European Statisticians, sixty-eighth plenary session, Geneva, 22–24 June 2020 (ECE/CES/2020/10). Available at https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2020/ECE_CES_2020_10-2005282_E.pdf.

7.3. EXAMPLES OF COMMUNICATING DATA INTEGRATION

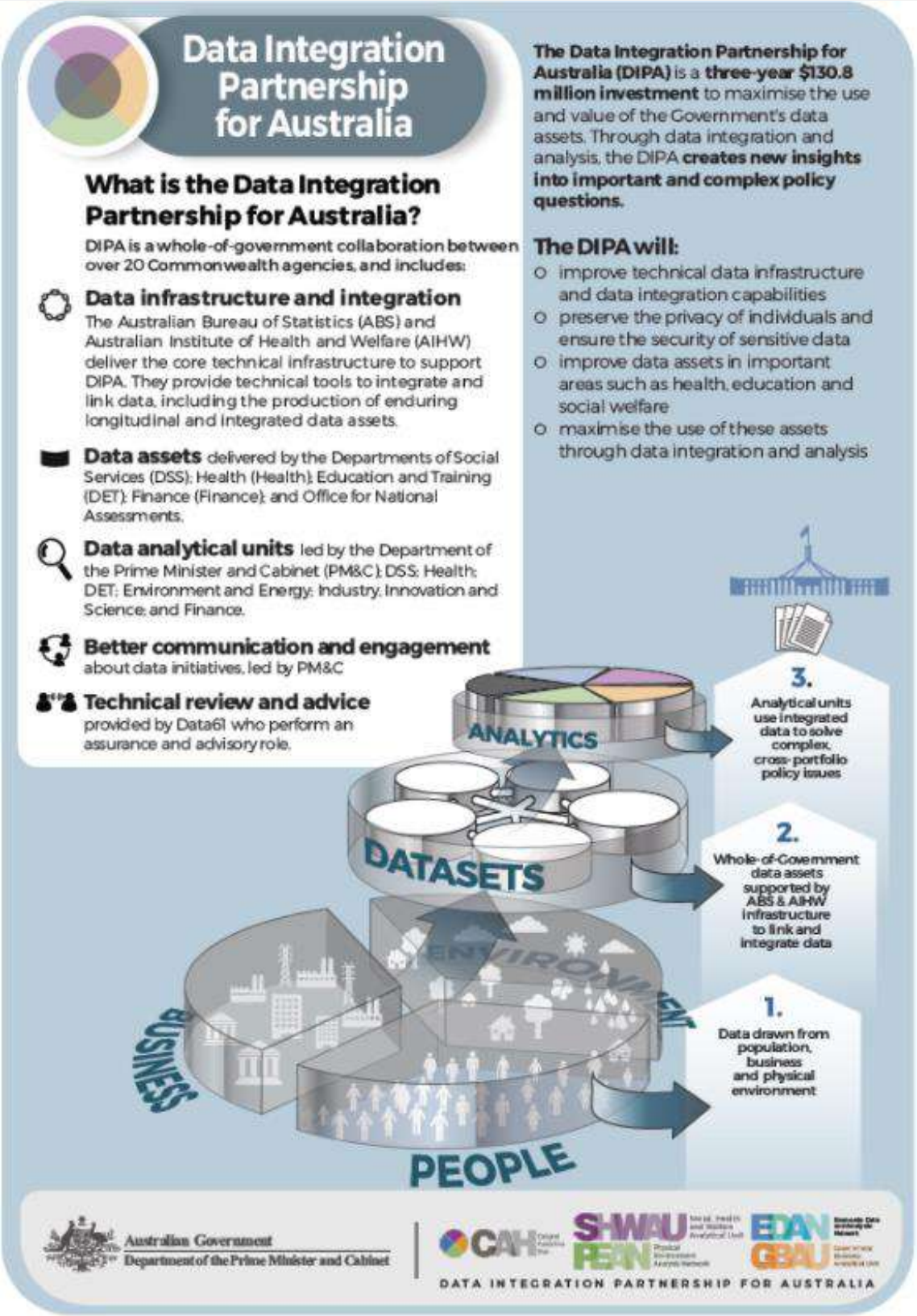
106. The Data Integration Partnership for Australia[66] has developed a diagram (figure 13) that provides a good example of communicating with a broad audience about data integration. The brochure has a few features worth noting:

- A 1, 2, 3 step diagram shows in a clear and simple way how data are sourced and transformed into datasets and analysis, and into meaningful information produced.
- Clear and brief explanation of the data integration partnership (main box with white background) − easy to understand.
- Clear explanation of the value proposition – the costs and the expected benefits.
- Branding government logos, giving status and credibility to the product.

107. Some ideas on communicating about data integration with different audiences were gathered in the ESCAP regional workshops on implementing data integration in Asia and the Pacific,[67] which are available on DI-CoP. These include suggestions on how to build relationships and mitigate risks to trust and credibility when communicating within the national statistical offices, with data suppliers, end users, citizens and respondents.

[66] See https://www.pmc.gov.au/public-data/data-integration-partnership-australia.
[67] See https://www.unescap.org/events/regional-workshops-implementing-data-integration-asia-and-pacific-round-1.

Figure 13. Data Integration Partnership for Australia diagram



Source: See https://www.pmc.gov.au/public-data/data-integration-partnership-australia.

Chapter 8

TYPES OF DATA INTEGRATION

8.1. TRADITIONAL DATA SOURCES (SURVEYS, CENSUSES, ADMINISTRATIVE SOURCES)

8.1.1. ADMINISTRATIVE SOURCES

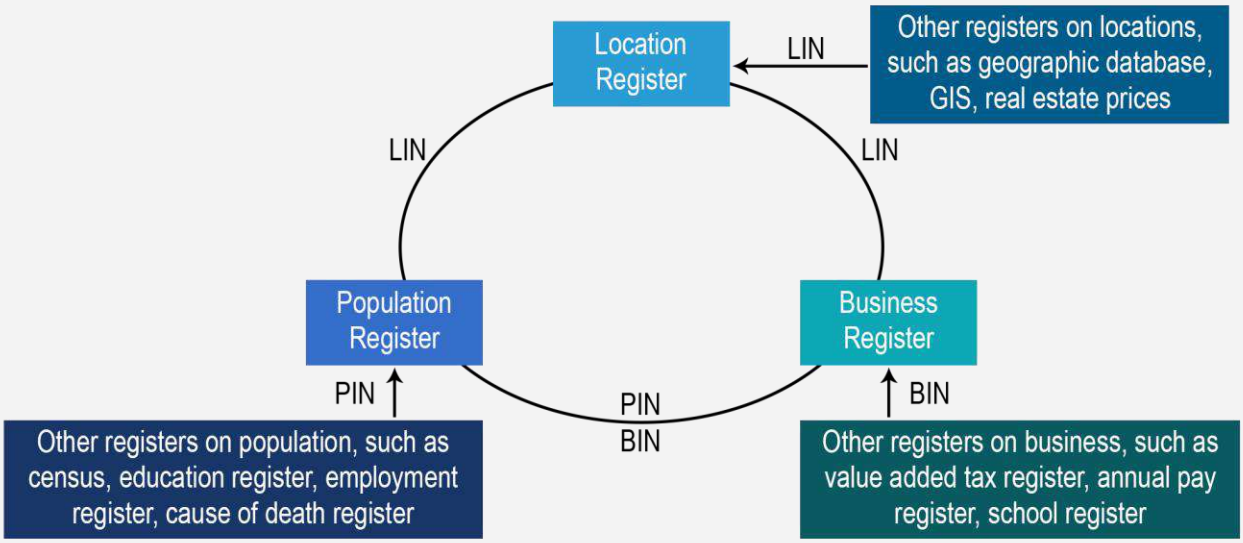
108. The subject of integrating administrative data is not new; for example, gross domestic product (GDP) is produced by integrating various statistics, such as production value, sales, exports and imports. Countries are using administrative data to produce official statistics to different and varying extents, depending on the level of development of their statistical systems and the quality of the administrative data available.

109. The most developed approach for using administrative data in producing official statistics is to build a system based on administrative data in which statistical registers are organized into a linked system. A *register* is defined as a systematic collection of unit-level data organized in way that updating is possible. Such systems were first developed in Nordic countries. This approach is thoroughly explained in the United Nations, ECE publication *Register-based Statistics in the Nordic Countries –- Review of Best Practices with Focus on Population and Social Statistics*.[68] Preconditions for this type of systems are a legal base, unified identification systems and cooperation among institutions.

110. A conceptual model of a statistical register system is illustrated in figure 14. In the figure, three base registers (population register, business register and location registers) are linked to each other and to other relevant administrative records using unique identifiers; namely PIN: personal identity number (such as national code, or social security code), BIN: business identity number, and LIN: location identity number (including geocode, a combination of address code and dwelling/building number, postal code, or any standard address of a location).

[68] *Register-based Statistics in the Nordic Countries –- Review of Best Practices with Focus on Population and Social Statistics* (United Nations publication, Sales No E 07.II.E.11). Available at https://unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

Figure 14. A conceptual model of a statistical register system,



Source: Adapted from Anders Wallgren and Britt Wallgren, *Register-Based Statistics: Statistical Methods for Administrative Data* (John T. Wiley and Sons, Inc., 2014).

111. The report of Usage of Administrative Data Sources for Statistical Purposes[69] indicates administrative data sources can effectively be used into two broad classifications:

- *Direct usage* includes situations in which there is an immediate link between the administrative data and statistical output. The administrative data may undergo various transformations, such as converting administrative units to statistical units (for example, for profiling businesses obtained from tax registers), or deriving statistical output variables from the unit's attributes, but, in essence, output is primarily sourced from the administrative data itself. Examples are direct tabulation (typically for full coverage of administrative data source) and substitution or supplementation for data collections, including whole and/or partial substitution for directly collected survey variables for subpopulations and/or variables of interest, or augmentation of directly collected survey variables.
- *Indirect usage* of administrative data describes situations in which administrative data play a supporting role in the creation of statistical output sourced primarily from either a survey or another administrative source. Examples are the use of administrative data in (a) the creation and/or maintenance of survey frames, such as in the development of statistical business registers (see User Guide for ADB Statistical Business Register),[70] (b) developing sampling designs by providing measures of variability for design variables and/or size measures, and facilitating sample selection, (c) editing and imputation by assisting in the construction of edit rule and/or imputation models, (d) indirect estimation and weighting by enabling the creation of population benchmarks, application of model-assisted or model-based frameworks, such as small area estimation (see the ADB publication on *Introduction to Small Area*

[69] See https://ec.europa.eu/eurostat/cros/system/files/Usage%20of%20Administrative%20Data%20Sources%20for%20Statistical%20Purposes.pdf.

[70] Asian Development Bank, *User Guide for ADB Statistical Business Register* (Manila, ADB, 2008). Available at https://www.adb.org/publications/adb-statistical-business-register-user-guide.

Estimation Techniques)[71] and addressing quality issues, such as non-response, and (e) validation of survey estimation and/or other administrative data source at micro- or macro-levels and assessing the quality of other administrative data source.

112. Administrative sources can provide full coverage of populations. This, however, depends on the population of interest and quality of administrative records. The potential of administrative data to cover whole populations can elevate the production of local area data to a level of detail not permitted by sample surveys, which is also an advantage in implementing local policies. Administrative sources may also make it possible to produce more frequent statistics. This depends on the nature of data and relevant administrative procedures. In some countries, the sources for administrative population registers, business registers, farmer registers and social security can be updated daily.

113. There are a number of challenges pertaining to the use and integration of administrative data. The quality of administrative datasets may be sufficient for administrative purposes but not for statistical purposes. Efforts to transform administrative datasets into statistical datasets often encounters quality and conceptual issues.

114. As administrative data are collected for non-statistical purposes, they are prone to conceptual inconsistencies. This can lead to a number of problems, including with coverage and bias. In some cases, such as with business statistics, administrative units do not necessarily correspond directly to the definition of the statistical units. This requires some modelling to convert the administrative units into statistical units. There may also be differences in the definitions of variables. It is important to have a thorough understanding of the impact of these differences. Sometimes, it is possible to influence the administrative definition by cooperating with the responsible authority. Administrative data are compliant with the laws in the country and in some instances, definitions and concepts are sufficient to provide national statistics but cannot be used in the context of international comparisons to other countries, where international or regional standards often apply.

115. Classifications is another important issue. Those used in administrative data may not be the same as the classifications used in statistical production. In cases of different classifications, the usual approach is to use correspondence tables and conversion tools based on additional variables that may be available for converting into more correct classification codes. Notably, however, even the same classifications may result in different data, especially when classifications are complex or the rules of a classification are difficult to apply. In administrative sources, coding is often carried out by the respondent, while a survey may have open questions and coding is often done by experts. Cooperation between the statistical organization and the administrative authority is an effective way to solve a part of the classification problem. The statistical organization can provide required training, assist in implementation and be responsible for maintenance of the classification. In some cases, statistical organizations have developed coding tools to be used by other organizations.

116. Missing data and errors also need to be considered. Missing data often occur because of a unit or variable non-response, but with regard to administrative sources, the causes can be different. It is important to identify if errors and missing data are systematic and

[71] Asian Development Bank, *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices* (Manila, ADB, May 2020). Available at https://www.adb.org/sites/default/files/publication/609476/small-area-estimation-guide-nsos.pdf.

then apply appropriate validation and editing rules. Inconsistency in the value of a single variable across different data sources should also be addressed.

117. Timeliness is another point in integrating administrative and survey data. Administrative data may not be available in time or may not coincide with the statistical reference period. Sometimes, such issues can be resolved by analysing the impact and if necessary, adjusting them using models.

118. When administrative data are used to supplement surveys or are integrated with other administrative data, it is desirable that the datasets contain common variables. The ideal situation is when datasets contain unique identifiers. If there are no unique identifiers, combinations of other available individual characteristics must be considered instead, such as name, address, date and place of birth, to identify identical subjects in different datasets.

119. Several guidelines, directives, standards and recommendations are available concerning administrative data. The following is an example of guidelines for dealing with administrative data that can be found on Statistics Canada webpages[72] (summarized to some extent):

- Maintain a continuing liaison with the provider of administrative records.
- Understand the context under which the administrative organization created the administrative programme, such as legislation, objectives and needs.
- Be cognizant that if the information provided to the administrative source can cause gains or losses to individuals or businesses, there may be biases in the information supplied, which can lead to unexpected coverage problems and biases.
- Collaborate with the designers of new or redesigned administrative systems.
- Develop an imputation or a weight-adjustment procedure to deal with non-response (unless non-respondents can be reached and responses obtained). Administrative sources are sometimes outdated. Accordingly, as part of the imputation process, give special attention to the identification of active and/or inactive units.
- Select the type of linkage methodology, such as record linkage or statistical matching, in accordance with the objectives. When the purpose is frame creation and maintenance, or data editing, record linkage of units should be used. In the case of imputation or weighting, record linkage should be used, but statistical matching may also be sufficient. When the sources are linked for performing some data analyses that are impossible otherwise, consider statistical matching, such as matching of records with similar statistical properties.
- To perform record linkage, make appropriate use of existing software.
- When data from more than one administrative source are combined, focus on reconciling potential differences in the concepts, definitions, reference dates, coverage, and the data quality standards applied at each data source.
- Some administrative data are longitudinal in nature, such as income tax, goods and services tax. When records from different reference periods are linked, they become very rich data mines for researchers. Remain especially vigilant when creating such longitudinal and person-oriented databases, as using them raises very serious privacy concerns.

[72] See https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/administrative-administratives-eng.htm#a2.

- Use identifiers with care, as a unit may change identifiers over time. Track such changes to ensure proper temporal data analysis. In some instances, the same unit may have two or more identifiers for the same reference period, which, in turn, introduces duplication in the administrative file. If this occurs, develop a mechanism to remove duplicates. A good relationship with the data provider and an understanding of the way that identifiers are managed is important.
- Document the nature and quality of the administrative data once assessed. Documentation helps statisticians decide the uses for which the administrative data are best suited. Choose appropriate methodologies based on administrative data and inform users of the methodology and data quality.

120. In response to the increasing importance of administrative data sources for producing official statistics, the United Nations Statistics Division and the Global Partnership for Sustainable Development Data have jointly convened the Administrative Data Collaborative,[73] a multi-stakeholder collaborative of national, regional and international agencies, aiming to strengthen the capacity of countries to use administrative data sources for statistical purposes. The collaborative addresses urgent and longer-term needs related to the access and use of administrative data for statistical purposes, building on advances made in various sectors and by different partners. The collaborative is a platform to share resources, tools, best practices and experiences. It is intended to contribute towards raising awareness among the members of national statistical systems about the benefits of sharing and combining administrative sources to enhance the quality, timeliness, coverage and level of disaggregation of statistical data.

8.1.2. SURVEYS AND CENSUSES

121. Unlike administrative data, integration of sample surveys data is not very common. Sample surveys cover a very small portion of a population, often lack unique identifiers and follow different sampling designs, reference time and target populations. Integrating or harmonizing survey processes makes it possible to generate integrated statistics relevant to subsets of a population. In addition, the potential to produce more comprehensive and frequent data and statistics from a survey provides good opportunities to combine survey data with censuses for production of small area estimations.

122. An integrated survey programme is intended to share concepts, definitions, classifications, sampling frames/scheme/tools, related materials, survey personnel and facilities across multiple surveys and survey rounds. Integration offers gains in efficiency and quality compared to when each survey is designed and carried out independently. Integrated survey programmes improve harmony and consistency of results, reduce respondent burden and cost of fieldwork operations, increase effective sample size for multi-purpose analyses, and foster small area estimations. Some of examples of integrated survey programmes are the United Kingdom Integrated Household Survey

[73] See https://www.data4sdgs.org/initiatives/administrative-data-collaborative.

programme,[74] the Rwanda Integrated Business Enterprise Survey,[75] the United States re-engineering the Census Bureau's annual economic survey [76] and Agriculture Integrated Survey programme (AGRISurvey) [77] of the Food and Agriculture Organization of the United Nations.

123. Two key aspects of integrated surveys are *common sampling design* and *common modular questionnaire*:

- Common sampling designs serve two main purposes: (a) enabling use of a master sample[78] for all surveys, which, in turn, increases cost-efficiency, enables controlling of respondent burden and promotes field operation and facilitates panels in the sample, and (b) providing a harmonized weighting system and facilitating multipurpose analysis by combining results from different surveys. Generally, sampling designs for integrated surveys are complex in order to provide enough flexibility for various selection procedures for the surveys being integrated.
- A common modular questionnaire is key in integrating surveys. It normally provides one core module, which is a minimum set of questions common across all surveys. The questionnaire is used to establish the link between integrated surveys, harmonize the key questions and reduce cost and respondent burden. The common questionnaire has different modules, each collecting data specific to one survey.

124. It is possible to use *statistical matching* to integrate data from harmonized surveys. As mentioned earlier, statistical matching involves the integration of data sources with usually distinct samples from the same target population in order to study and provide information on the relationship of variables not jointly observed in the datasets. Matching procedures can be regarded as imputation of the target variables from a donor survey to a recipient survey (see chapter 6).

125. The most widely used techniques for integrating survey and census data are small area estimation (SAE) methods. Survey data provide detailed information about a target population, some of which cannot be collected in censuses, such as income, expenditure, nutrition and mortality. At the same time, sample surveys collect data only on a small portion of a population, often selected randomly following a sampling scheme. It is, therefore, not possible to produce survey estimates for very small subpopulation groups or small geographic areas with desirable accuracy. SAE methods apply statistical models and rely on the variables present in both a survey and census to borrow strength from comprehensive coverage of census data and estimate survey variables in small population groups. SAE applications, techniques and data sources are very diverse. A

---

[74] United Kingdom, Office for National Statistics, "Integrated Household Survey", 3 February 2016. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/sexuality/methodologies/integratedhouseholdsurvey.

[75] Department of Economic Statistics, National Institute of Statistics of Rwanda, "Integrated Business Survey 2018" (December 2019). Available at https://www.statistics.gov.rw/publication/integrated-business-enterprise-survey-ibes-2018.

[76] National Academies of Sciences, Engineering, and Medicine, *Reengineering the Census Bureau's Annual Economic Surveys* (Washington, D.C., The National Academies Press, 2018). Available at https://www.nap.edu/read/25098/chapter/10.

[77] Food and Agriculture Organization of the United Nations, "The Agriculture Integrated Survey programme −AGRIsurvey". Available at http://www.fao.org/3/ca6785en/ca6785en.pdf.

[78] See United Nations Department of Economic and Social Affairs, Statistics Division, "Designing Household survey samples: practical guidelines", Studies in Methods, series F. No. 98. (New York, United Nations, 2005). Available at https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf.

well-known method is ELL (Elbers, Lanjouw and Lanjouw),[79] developed by the World Bank, which is widely applied for poverty mapping[80] (a software[81] has also been developed to implement this method). The recent publication of ADB, *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices* provides a good introduction to some SAE techniques and application of R software for their implementation.

126. In general, SAE refers to any of the several techniques that entails combining the survey data with other types of auxiliary data, such as administrative data and census information, which have wider coverage, in order to enhance the survey estimator and provide more reliable granular statistics, without increasing survey costs. Figure 15 shows the major processes in carrying out SAE. With increasing access to new data sources and enhancing computation power, the techniques and applications of SAE are also evolving. For instance, application of spatial information and tools in SAE has increased in recent years, both as a source of data and for visualization purposes.[82] In addition, application of big data for SAE has emerged with promising results. For instance, satellite imagery and nightlight data are applied to produce poverty estimates in very small areas, such as mapping poverty through data integration and artificial intelligence, as explained in two ADB publications.[83,84]

127. The Joint Intersecretariat Working Group on Household Surveys and Interagency Expert Group on Sustainable Development Goals (IAEG-SDGs) Task Force on Small Area Estimates led by the United Nations Statistics Division and comprised of members of national statistical offices,[85] international agencies, academia, and non-governmental agencies are developing a toolkit on SAE for the Sustainable Development Goals. The toolkit will include an overview of small area estimation methods, step-by-step guides on using small area estimation methods, country examples and case studies, visualization and communication of small area estimates, and software packages for SAE. The tentative title of the toolkit is “SAE4SDG”.

128. The Asian Development Bank, in partnership with United Nations Statistics Division, prepared a draft practical guidebook on data disaggregation for SDGs, which was presented as one of the background documents at Fifty-second session of the United Nations Statistical Commission, held virtually from 1 to 3 March 2021. The guidebook

---

[79] Chris Elbers, Jean O. Lanjouw and Peter Lanjouw, “Micro-level estimation of poverty and inequality”, *Econometrica*, vol. 71, No. 1 (January 2003) pp. 355−364. Available at https://are.berkeley.edu/~ligon/Teaching/ARE251/elbers-etal03.pdf.

[80] Tara Bedi, Aline Coudouel and Kenneth Simler (eds.), *Using Poverty Maps to Design Better Policies and Interventions* (Washington, D.C., the International Bank for Reconstruction and Development/World Bank, 2007). Available at https://documents.worldbank.org/en/publication/documents-reports/documentdetail/875491468320356640/more-than-a-pretty-picture-using-poverty-maps-to-design-better-policies-and-interventions.

[81] See https://www.worldbank.org/en/research/brief/software-for-poverty-mapping.

[82] See https://www.worldpop.org/.

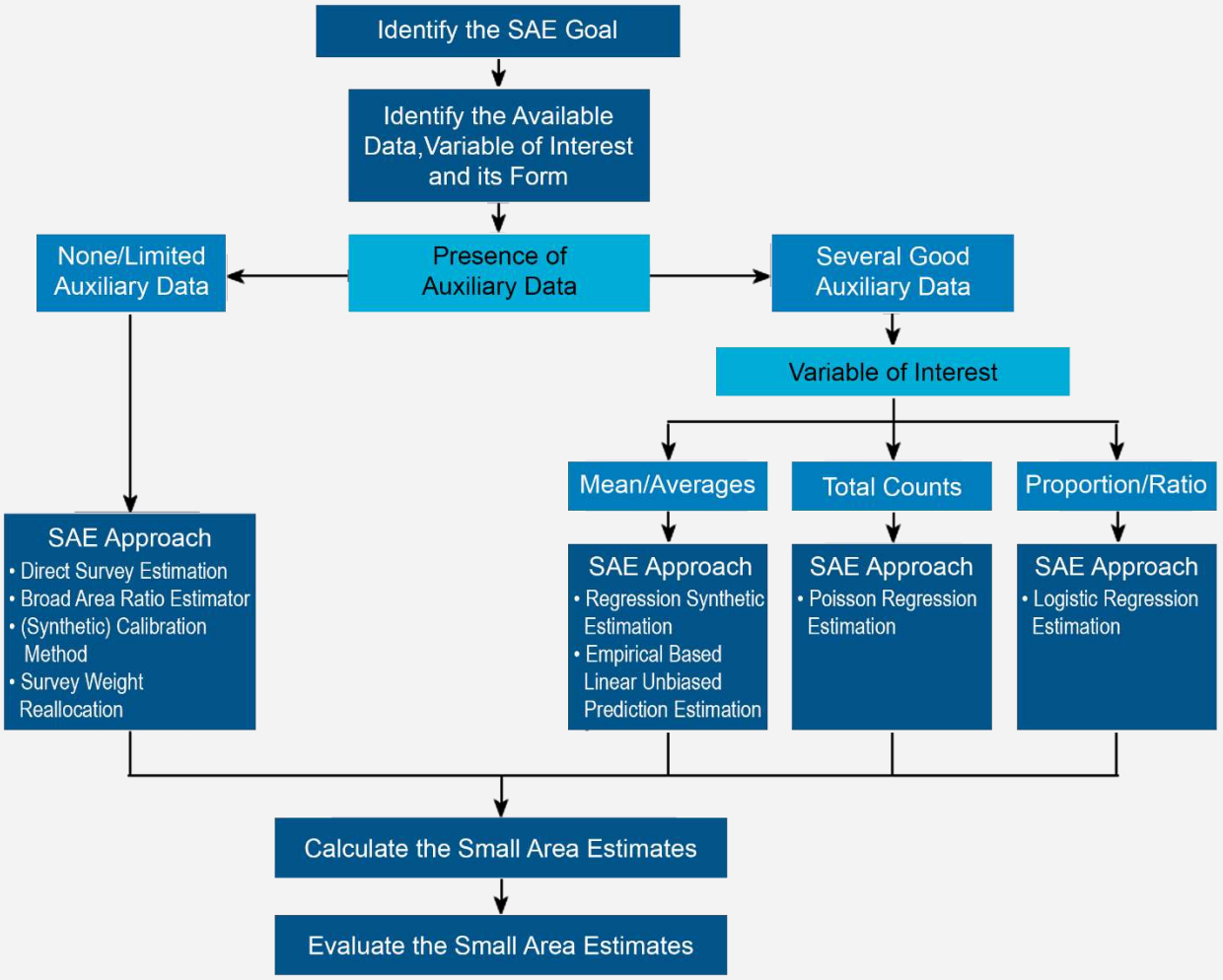
[83] See Asian Development Bank, *Mapping Poverty through Data Integration and Artificial Intelligence* (Manila, ADB, 2020). Available at https://www.adb.org/sites/default/files/publication/630406/mapping-poverty-ki2020-supplement.pdf.

[84] Asian Development Bank, *Mapping the Spatial Distribution of Poverty Using Satellite Imagery in the Philippines* (Manila, ADB, 2021). Available at https://www.adb.org/sites/default/files/publication/682851/mapping-poverty-satellite-imagery-philippines.pdf.

[85] See https://unstats.un.org/iswghs/task-forces/task-forces-round2/.

provides background materials on issues and experiences of countries regarding data disaggregation for the Sustainable Development Goals.[86]

Figure 15. Major processes in undertaking small area estimation



***Source**:* Australian Bureau of Statistics. "A guide to small area estimation" - version 1.1, Internal ABS document (2006).

Note: SAE, small area estimation.

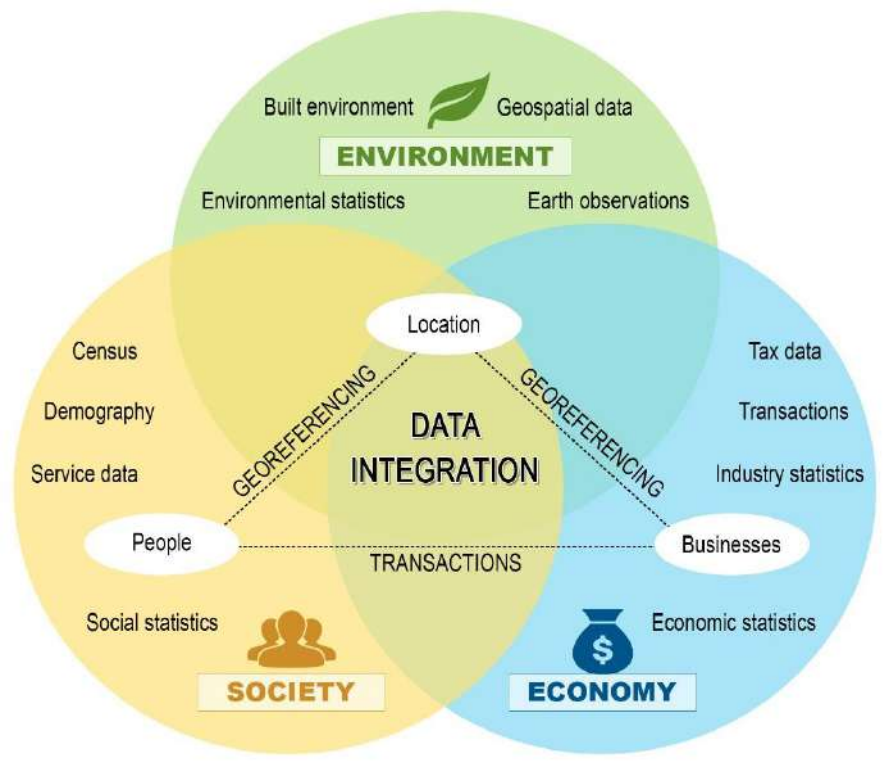
[86] See https://unstats.un.org/unsd/statcom/52nd-session/documents/BG-3a-Practical_guidebook_on_data_disaggregation_for_the_SDGs-E.pdf.

8.2. NEW SOURCES OF DATA

8.2.1. GEOSPATIAL INFORMATION

129. Statistical data are almost always related to a certain physical space, such as a municipality, state, country and region. Each level is useful for different actors and different types of decisions. Many of those decisions are influenced by the physical characteristics of a location, and each decision can have an impact on the environment. Amounts of natural resources, soil types, weather conditions, communications infrastructure and facilities are examples of geographic information that is indispensable in order to fully understand the figures generated through official statistics. Linking data on people and businesses to a place or geographic location, and integrating them with geospatial information through the medium of location can help to better understand social, economic and environmental issues to a much greater degree than when viewing statistical or geospatial information in isolation (figure 16).

Figure 16. Location as a link between society, the economy and the environment



Source: United Nations, Department of Economic and Social Affairs, Statistics Division, "The Global Statistical Geospatial Framework" (2019).[87]

[87] Available at http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf.

130. The Working Group on Geospatial Information[88] had identified that integration of statistical data and geospatial information was key for the production of a number of Sustainable Development Goals indicators (15 indicators in which geospatial information has a direct contribution and nine indicators in which geospatial information has a significant or supporting contribution). The ESCAP Stats Brief entitled "Geospatial information and the 2030 Agenda for Sustainable Development"[89] includes country examples of Sustainable Development Goals indicators compiled using Earth observations and geospatial data.

131. The geospatial and statistical data integration landscape is complex. The Global Statistical Geospatial Framework (GSGF)[90] is vital for a consistent and systematic approach to linking geospatial and statistical data. The GSGF is a common method for geospatially enabling statistical and administrative data to ensure that data from a range of sources can be integrated based on location and also can be integrated with other geospatial information.

132. The Global Statistical Geospatial Framework (figure 17) is a high-level framework that facilitates consistent production and integration approaches for geostatistical information. It is generic and permits application of the framework principles to the local circumstance of individual countries. Fundamentally, GSGF enables the following:

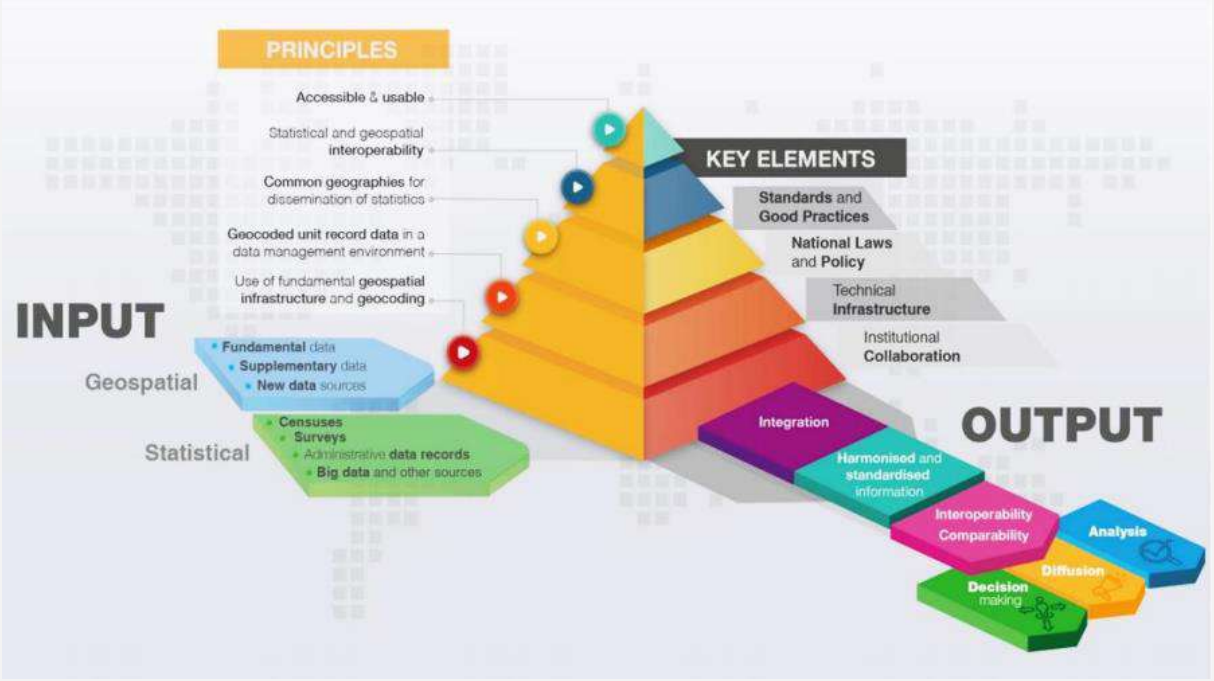
- Integration of data to support the measuring and monitoring of the targets and global indicator framework for the Sustainable Development Goals of the 2030 Agenda for Sustainable Development and the 2020 Round of Population and Housing Censuses;
- Comparisons at local, subnational, national, regional and global levels for decision-making processes within and between countries and thematic domains;
- Data sharing between institutions through interoperability of geospatial and statistical information and the development of common tools and applications;
- Unlocking of new insights and data relationships that is not possible by analysing socioeconomic, environmental or geospatial data in isolation;
- Increased information on smaller geographical areas;
- Increased awareness of methods and tools to assess and manage disclosure risks and to enhance privacy in the collection, storage and dissemination of information;
- Conditions for investment and capability-building in geospatial and statistical information;
- Integration of new sources of data to inform the production of high-quality geospatial information, for example Earth observations and other complementary data sources;
- Strengthening of institutional collaboration between the geospatial and statistical communities.

[88] See United Nations, Department of Economic and Social Affairs, Statistics Division, "Working Group on Geospatial Information" (July 2019). Available at http://ggim.un.org/UNGGIM-wg6/.

[89] United Nations, Economic and Social Commission for Asia and the Pacific, "Geospatial information and the 2030 Agenda for Sustainable Development", Stats Brief, Issue No. 27 (December 2020). Available at https://www.unescap.org/sites/default/d8files/knowledge-products/Stats_Brief_Issue27_Dec2020_Geospatial_data_for_SDGs.pdf.

[90] United Nations, Department of Economic and Social Affairs, Statistics Division, "The Global Statistical Geospatial Framework" (2019). Available at http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf.

Figure 17. The Global Statistical Geospatial Framework − Inputs through to outputs



Source: United Nations, Department of Economic and Social Affairs, Statistics Division, "The Global Statistical Geospatial Framework" (2019).

133. The GEOSTAT project[91] involves cooperation between Eurostat and the European Forum for Geography and Statistics with the objective to promote grid-based statistics and more generally to work towards the integration of statistical and geospatial information in a common information infrastructure. The aim of the GEOSTAT projects is to develop common guidelines for the collection and production of spatial- and grid-statistics within the European Statistical System.

134. Integration can take place at any stage of the statistical production process, as outlined in GSBPM. The integration entails geocoding of statistics, spatial analysis, and creating statistical maps. As part of the integration process, the following steps may be carried out:

- Geocoding statistical information at unit-record level;
- Processing and manipulation of statistical information using spatial analysis techniques with the purpose of selecting information or deriving new information with a focus on their spatial characteristics, such as buffering around spatial features;
- Supporting a more efficient and flexible statistical production process with geospatial information, such as for surveying and sampling or a field operation;
- Combining statistical end products with geospatial information in statistical maps;

[91] See https://www.efgs.info/geostat/.

- Improving the quality of existing statistical products adopting spatial models, such as commuting information by calculating journey times based on detailed transport networks.

135. All statistical phenomena that can be associated to a location are in principle relevant for the integration of statistical and geospatial information. Location in this context means the location of the most individual observation at a unit record level. In most cases the location is a point with coordinates or a precise address. However, other spatial reference frameworks, such as lines or polygons, are relevant in addition to representing road segments or areas with a certain land cover.

136. Integration of geographical data with statistical data is intended to improve the value of the statistical information that is being produced. Geographic information system (GIS) should be used as much as possible at all stages (inventory, preparation, progress, monitoring, dissemination of results) of the geospatial integration. Wherever it is possible, data should be collected with reference to an address point; the results can then be disseminated using any desired spatial divisions. GIS technology should be considered only at a level appropriate to the skills and resources available and constitute an integral part of the overall work of a national statistical office.

137. Some of the opportunities arising from integration of geospatial data with statistical information are the following:

- Increasing added value of statistical and spatial data;
- Improving usability of statistical data for evidence-based decision-making at different geographical levels, such as municipality, state and country:
- Improving visualization of statistical information;
- Enhancing data and statistics services;
- Strengthening statistical analysis by including geographical and environmental elements;
- Facilitating integration of different data sources, such as administrative data and mobile data;
- Improving flexibility to use statistical information by external users, such as for scientific, environmental and humanitarian purposes and for rescue operations;
- Enhanced collaboration between mapping agencies with statistical institutions.

138. Some of the challenges in integrating geospatial data with statistical information are the complexity of the process, access to data, interoperability, budget restrictions, legal issues and confidentiality. To tackle these challenges, it is necessary to apply a consistent and systematic approach, enhance skills, expertise and knowledge in new areas, prepare required technology and infrastructure and build and improve collaboration with mapping agencies and securing access to data.

139. Governments in the region have been actively developing national geospatial frameworks. For example, India developed a multi-layer GIS platform, Bharat Maps,[92] which integrated satellite imagery from multiple agencies. The Government of Indonesia initiated the One Data and One Map Policy.[93] Mongolia developed an integrated

[92] See https://bharatmaps.gov.in/.

[93] Land portal, "Indonesia launches One Map Policy Geoportal to improve investment climate", 11 December 2018. Available at https://landportal.org/node/77328.

geospatial information framework and is reviewing its spatial data infrastructure law (Mongolia: National Geospatial Infrastructure).[94]

140. While the national geospatial frameworks guide the implementation process, they do not necessarily address the issue of integrating it with the statistics framework. The Australian Bureau of Statistics tackled the challenge of integration of geospatial and statistical information for an evidence-based decision-making by developing the Statistical Spatial Framework.[95] which was adapted for GSGF. Countries are building their capacity for implementing GSGF through taking various steps to facilitate the execution, baselining, and assessment at the country-level plans. In addition to focusing on building the formal fundamentals, several countries are implementing programmes using geospatial data and Earth observation data for various needs. Many countries in the region are using geospatial data with a specific focus on assessing the progress made towards achieving the Sustainable Development Goals. Some examples of this are provided in section 8.2.2.

141. With the support of the United Nations Global Geospatial Information Management for Asia and the Pacific (UN-GGIM-AP), a project, which aims to connect geospatial data sources across the Asia-Pacific region and aggregate them onto an interoperable regional geospatial data platform over the Sustainable Development Goals "Decade of Action" through to 2030 was proposed in 2020.

142. Under the overarching framework of the Asia-Pacific Plan of Action on Space Applications for Sustainable Development (2018–2030), the platform is intended to provide data services across the Plan's six thematic areas, namely disaster risk reduction (drought and floods); natural resource management (land and water); connectivity (city/urban); social development (health and pandemics); energy (renewable energy); and climate change (environment and air quality). The platform will promote more open, interoperable and orderly sharing of geospatial data between space-faring data-supply countries and regional data-users in the Asia-Pacific region. This, in turn will deepen understanding of complex sustainable development challenges and promote solutions for implementation of the 2030 Agenda for Sustainable Development.

143. Initial key activities in the first phase of the Plan (2020–2022), funded by the Republic of Korea, include an analytical and conceptual study to identify good models and practices of sharing geospatial data across existing geo-data platforms and the development of a set of foundation documents on such topics as objectives, governance, finance, structure, activities and operational principles for intercountry sharing of data, as well as a work/budget plan for 2023–2026.

144. The Economic and Social Commission for Asia and the Pacific, in its role as the secretariat of UN-GGIM-AP, will develop the above documents in full consultation with member countries of UN-GGIM-AP and base them on the conceptual research. The consultations will focus on developing a plan and actions.

---

[94] Enkhtur Bayarmaa, "Mongolia: national geospatial data infrastructure", UN-GGIM-AP WG-3 (October 06, 2020). Available at https://un-ggim-ap.org/sites/default/files/media/Working%20Groups/WG3/WG3%20Webinar_2020-10-07/Presentations/3.%20Mongolia.pdf.

[95] See https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material.

8.2.2. BIG DATA

145. The past several years, interest in integrating new data sources, such as big data, into the production of official statistics among national statistical offices has been increasing. Consequently, experimentation and integration efforts have been multiplying across the region.

146. Commonly, "big data" is referred to as exhibiting certain characteristics in the form of multiple Vs (volume, velocity, variety, variability, veracity, validity, vulnerability, volatility, visualization and value). This list has been growing and may include additional characteristics, as the technologies and data types they generate continue to evolve. Most of these data sources are captured and owned by the private sector, a development facilitated by cloud technology. As governments also tend to collect data with several of the mentioned characteristics, certain types of administrative data may also be considered as "big data".

147. The Economic Commission for Europe classifies big data into three main categories:

- *Social networks* (human-sourced information) – digital recordings of human activity, such as data from social platforms, videos, the Internet searched and mobile content;
- *Traditional business systems* (process-mediated data) – recorded events and transactions by the private and public sectors, such as medical records, commercial transactions and e-commerce;
- *Internet of things* (machine-generated data) – data derived from sensors, such as mobile signalling data, satellite imageries, weather/pollution sensors, traffic sensors and web logs.

148. The characteristics of big data sources, such as the Vs, and the fluid structure of some require new technologies and methods to process, store, and analyse them. Machine learning, artificial intelligence and deep learning are examples of new approaches being actively deployed by national statistical offices to explore big data sources. Examples of governmental agencies and countries using these techniques are the Department of Statistics Malaysia, which is exploring big data analytics (see STATSBDA Portal[96]) and the Philippines and Thailand, which are working with ADB on mapping poverty through data integration and artificial intelligence.[97]

149. Big data sources and methods differ from sample surveys in multiple ways. Because big data are often collected continuously, dynamic rather than periodical monitoring is possible. Full area coverage, relevant in the case of remote sensing, can provide information about the entire area of interest, rather than a just a sample. This may not hold true in the case of data from social networks or traditional business systems, depending on the country's level of digital development. For example, using social media data to capture sentiments will only attain the sentiments of people who use social media, resulting in significant bias. The multidimensionality of big data in which a multidimensional model views data in the form of a data cube with multiple dimensions, rather than a tabular representation, increases the granularity of data at the cost of complexity. The timeliness of big data can sharpen the relevance and precision of the

[96] See https://statsbda.dosm.gov.my/.

[97] Arturo M. Martinez Jr., "Mapping poverty through data integration and artificial intelligence", presentation. Available at https://www.unescap.org/sites/default/files/Session4_ADB_mapping_poverty_DI-CoP_WS.pdf

monitoring tools, raising their importance in urgent situations, such as when there is no time or funding to run a large survey. The additional properties of big data vis-à-vis the traditionally collected data are matched by new technologies that allow for analysis and visualization of data, previously unavailable. Accordingly, to make use of big data sources for official statistics and to remain a relevant player in providing useful statistics to the government in the era of big data, statistical organizations must upgrade their internal human and technological capacities.

150. Although these data sources are collectively referred to as “big data”, each type requires different technologies, skills, and methods to process and analyse them; they come in different types and from different sources. Accordingly, individual data types and data sources should be analysed in the context of individual countries and their specificities in terms of their legal framework, technological infrastructure and Internet and digital service penetration to estimate the data source fit for integration into the official statistics in that country. For example, in the context of statistical compilation, social media data or credit card transaction data may provide biased results in a country with a limited social media or low banking penetration.

151. When considering big data integration into official statistics, national statistical communities need to address the following challenges: legislative; privacy; financial, management; methodological; and technological (see In-depth review of big data).[98]

152. When using big data, national statistical offices should be cautious because of the following reasons. First, there is need to investigate the relevance of data coming from these new sources. With regard to quality, big data may be subject to significant coverage issues (being not representative of the target statistical population), and/or measurement errors (particularly in the case of human-sourced information). Second, the compatibility of concepts and definitions should be checked. Shortcomings of the data being used to produce official statistics should be properly understood, stated and addressed (through pre-processing phases). While big data can be used to improve the efficiency of survey estimates as described in The ABC of Big Data,[99] replacement of surveys with big data should be considered with circumspection.

153. In pursuing big data integration, statistical organizations are likely to incur financial costs on two fronts: data *acquisition*, particularly if there is no legislation covering acquisition of external data; and *big data infrastructure*, which is necessary to support the use of big data. While costs may be significant, they should be compared against those related to traditional data collection; the result may show efficiency gains across the government through access to timely and granular information with a significant lower response burden.

154. As big data require new tools for data capturing, processing and integration, statistical organizations may need to develop the right big data IT infrastructure, or partner with another organizations, in accordance with the national legislation. Some countries that practice “data nationalism”, require homegrown data to reside physically in their country. This may limit the choice of cloud service providers for public agencies, obliging them to

[98] United Nations, Economic Commission for Europe, “In-depth review of big data”, Conference of European Statisticians, sixty session plenary session, Paris, 9−11 April 2014 (ECE/CES/2014/7). Available at https://unece.org/DAM/stats/documents/ece/ces/2014/7-In-depth_review_of_big_data.pdf.

[99] Siu Ming Tam, “The ABC of big data”, Data Integration Workshop, 26 November−3 December 2020. Available at https://www.unescap.org/sites/default/files/Session3_ABC_of_Big_data_DI-CoP_WS.pdf.

either develop their own internal IT infrastructure or collaborate with national cloud service providers.

155. Technical capacity and big data specialists are crucial for the success of big data integration into the production of statistics. While some national statistical offices develop internal capacities, others choose to partner with academia or specialized agencies, such as national space agencies for Earth observation data processing expertise. In some instances, particular areas of statistics, such as the generation of CPI through web scraped online price data or scanner data, have been outsourced to private companies, which obtain the data, process them and share results with the national statistical offices. While such examples exist, a number of national statistical offices are gradually building their own in-house capacity.

156. Access to big data sources and types of data partnerships with the private sector data holders vary across countries. Some statistical organizations negotiate and access data directly from data holders, while others collaborate with governmental ministries and national telecommunications regulators that regulate data access conditions, as in the case for accessing mobile network operator data in Georgia and Viet Nam.

157. In the absence of a relevant regulatory framework guiding data access from the private sector, companies may be reticent to share data, especially when it contains personal information, such as mobile phone data or financial transaction data, regardless of the level of aggregation. Instead, when in a partnership with the national statistical office, they may opt to share the outputs obtained through methods and algorithms developed by the statistical organization, shifting access to data services. Data access may also be subject to costs. In the absence of legislation regulating financial modalities of external data access, some statistical organizations obtain data on a commercial basis. Other organizations may obtain access to data without costs, especially when access is guided and facilitated by the regulator. Other models of data access may involve outsourcing to private companies that cover the entire process from data acquisition to statistics generation, or to subsidiaries of the statistical organization, as exemplified by Data Ventures New Zealand.[100]

158. Some statistical organizations seek or gain access to administrative big data that are collected from the private sector, as in the case of tax data. The national statistical offices of Mongolia, New Zealand and the Russian Federation are either using or exploring this data source. Depending on the national legislation that guides data collection, data sharing, and statistics production, data may not be handled or appropriated for uses beyond what is initially intended. This, in turn, may require legal changes to data processing and sharing within the government or to data privacy and sharing consent at the point of collection.

159. Big data pilot initiatives differ from efforts to integrate big data into the production of official statistics. For pilot projects, access may be obtained for a sample, while integration of private sector data into official statistics may require full data access (under models described above) and different types of agreements with data providers. Most of the projects in the region are still in the piloting or experimenting phase, with very few moving towards replacement of traditional data collection and statistics production. Some of the big data projects, however, do not extend beyond the pilot phase. The reasons behind this vary from unsatisfactory data quality to lack of technical or financial resources required for big data integration. Nevertheless, the pilot initiatives provide valuable

[100] See https://dataventures.nz/.

insights into the potentials and shortcomings of the data sources and the required resources.

160. There are also risks associated with the use of big data. Among these risks are data continuity, which may be subject to an agreement between the data provider and the statistical organization, company's life on the market and the continuity of financial resources of the statistical organization for external data access. Another risk associated with online data, such as online price data, is the potential change in data structure and the classification or access path without notice. Data privacy should be preserved, as any compromise of this can raise image and reputational issues. Public trust is another aspect that needs to be considered when exploring private or sensitive data sources. It can be nurtured through awareness-raising, communication and transparency on the data sources used and how they are used.

161. The following are examples of the most explored big data sources by statistical organizations in the Asia-Pacific region:

- ***Online price data*** (web scraped) and ***scanner data*** for compilation of CPI. Scanner data require partnerships with retailers. Only a few national statistical offices in the region, namely the Australian Bureau of Statistics, Statistics New Zealand, Japan Statistics Bureau and the National Statistics Office of Georgia, have secured access to scanner data directly or through third-party data providers.[101] Another example is from the State Statistical Committee of the Republic of Azerbaijan,[102] which obtains data directly from electronic databases of trade networks (different from scanner data as it does not include information on the quantities of goods sold). Statistics Indonesia,[103] the Department of Statistics Malaysia,[104] the Department of Statistics Singapore and the General Statistics Office of Viet Nam explore online price data through web scraping and web scrawling techniques.

- ***Mobile network operator data***, such as signalling data and call detail records for producing mobility-related indicators, for example migration and tourism statistics, and population estimates. National statistics offices that use mobile network operator data are Statistics Korea, Statistics Indonesia, [105] National Statistics Office of

[101] United Nations, Economic and Social Commission for Asia and the Pacific, "Incorporating non-traditional data sources into official statistics: the case of consumer price indexes", Available at https://www.unescap.org/sites/default/d8files/knowledge-products/incorporating_non_traditional_sources_CPI.pdf.

[102] State Statistical Committee of the Republic of Azerbaijan, "Price statistics: application of a new approach to data collection". Available at https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2020/vebinar_Aze.pdf.

[103] Alifa Putri Wijaya, "Study of consumer price index based on e-commerce in Indonesia", presentation for APES week 2019. Available at https://communities.unescap.org/system/files/paper_58_indonesia_e-commerce_cpi.pdf.

[104] Mazliana Mustapa, "Leveraging online price data from web crawling", presentation, Asia-Pacific Stats Café, 30 November 2020. Available at https://www.unescap.org/sites/default/files/Leveraging_online_price_data_from_web_crawling_Malaysia_Stats_Cafe_30Nov2020.pdf.

[105] Statistics Indonesia, "The use of mobile positioning data; Indonesia's experiences", presentation for the International Symposium on the use of big data for official statistics, Hongzhou, China, 16−18 October 2019. Available at https://unstats.un.org/bigdata/events/2019/hangzhou/presentations/day2/6.%20The%20Use%20of%20Mobile%20Positioning%20Data%20Indonesia%E2%80%99s%20Experiences.pdf.

Georgia, National Bureau of Statistics of China, and Data Ventures of Statistics New Zealand.[106]

- ***Earth observation*** data collected through UAV (drones) or satellite imagery in the area of environment (land cover and land use mapping) and agriculture (crop monitoring) statistics or for poverty estimates. Several national statistical offices in the region collaborate with national space agencies, relying on the latter's data and technical experience in the production of Earth observation-derived statistics. As many other specialized agencies and ministries are involved in the development of environmental and agricultural statistics, they are also developing the Earth observation-related competencies and integrating Earth observation data into statistics production. Among the national statistical offices that use Earth observation data are the Australian Bureau of Statistics, which is collaborating with Geoscience Australia to develop environmental and agricultural statistics and tools, the Ministry of Statistics and Programme Implementation of India, which is collaborating with the National Remote Sensing Centre in the production of environmental accounts[107] and the Department of Statistics Malaysia, which is collaborating with the Malaysian Space Agency to identify living quarters in remote and inaccessible areas. In addition, ADB has supported several governments in South-East Asia in using Earth observation data to produce agriculture statistics[108] and to map poverty,[109] and ESCAP has published a guide on producing land cover change maps and statistics[110] to support countries in using QGIS, a free and open-source cross-platform geographic information system, to generate maps in accordance with the System of Environmental Economic Accounting.

- ***Social media*** data for conducting sentiment analysis. Statistics Indonesia,[111] the General Statistics Office of Viet Nam,[112] Statistics Korea and Statistics New Zealand are using social media to measure economic sentiment.

- Other data sources, such as financial and credit card transactions, and smart meter data are being explored to a smaller extent.

---

[106] See https://dataventures.nz/index.html.

[107] India, Ministry of Statistics and Programme Implementation, EnviStats India 2010. Available at http://www.mospi.nic.in/sites/default/files/reports_and_publication/statistical_publication/EnviStats2/ES2_2020_Complete_revised%20on%204_11_2020.pdf.

[108] Lea Rotairo and others, *Use of Remote Sensing to Estimate Paddy Area and Production: A Handbook* (Manila, ADB, 2019). Available at https://www.adb.org/sites/default/files/publication/496976/remote-sensing-paddy-area-production-handbook.pdf.

[109] See Asian Development Bank, *Mapping Poverty through Data Integration and Artificial Intelligence* (Manila, ADB, 2020). Available at https://www.adb.org/sites/default/files/publication/630406/mapping-poverty-ki2020-supplement.pdf.

[110] United Nations, Economic and Social Commission for Asia and the Pacific, "Producing land cover change maps and statistics: step by step guide on the use of QGIS and RStudio", 5 October 2020. Available at https://www.unescap.org/resources/producing-land-cover-change-maps-and-statistics-step-step-guide-use-qgis-and-rstudio.

[111] Asita Sekar Asri and Siti Mariyah, "Subjective Happiness Index based on Twitter in Indonesia". Available at https://communities.unescap.org/system/files/49._subjective_happiness_index_based_on_twitter_in_indonesia.pdf.

[112] Nguyen The Hung, "Listening the public opinion? An approach from big data with the case of revision GDP in the period 2010-2017 in Vietnam" (2020). Available at https://www.unescap.org/sites/default/files/APS2020/35_Listening_the_public_opinion-big_data_revision_GDP_2010-2017_GSO_Viet_Nam.pdf.

The ESCAP has mapped experiences of the national statistical offices in using big data sources to produce economic, population and social, as well as environmental and agricultural statistics. [113]

162. Big data sources can also be used in the compilation of Sustainable Development Goals indicators. Several countries in the region are conducting research on the big data opportunities for these indicators. For example, the Chinese Academy of Sciences has explored the potential of Earth observation data through piloting it in the compilation of 12 Sustainable Development Goals indicators,[114] while Deqing Province in China has integrated geospatial data into 14 indicators for reporting on the Sustainable Development Goals. [115] The Philippines Statistics Authority has explored citizen-generated data for Sustainable Development Goals reporting on 81 indicators in collaboration with civil society organizations, non-governmental organizations, development partners, member of the private sector, and academia. [116] The Data Research Center for Sustainable Development Goals research under the Statistics Research Institute of the Republic of Korea is conducting research on the use of innovative techniques and big data sources, such as GIS information, satellite imagery and public service transit data to measure Sustainable Development Goals indicators. The ESCAP has conducted a research on the use of big data sources for compilation of Sustainable Development Goals indicators highlighting country examples, focusing on countries in Asia and the Pacific. [117]

163. Big data remains a complementary data source to traditional data, in particular in areas where important data gaps persist. However, as data and technologies are evolving and their potential and limitations are researched, new data sources and big data methods may replace some conventional data collection and alter the way some of the official statistics are produced.

---

[113] For more information on country examples, see:
a. Economic and Social Commission for Asia and the Pacific, “Big data for economic statistics”, Stats Brief, Issue no. 28 (March 2021). Available at https://www.unescap.org/sites/default/d8files/knowledge-products/Stats_Brief_Issue28_Big_data_for_economic_statistics_Mar2021.pdf.
b. Economic and Social Commission for Asia and the Pacific, “Big data for population and social statistics”, Stats Brief, Issue no. 29 (April 2021). Available at https://www.unescap.org/sites/default/d8files/knowledge-products/Stats_Brief_Issue29_Big_data_for_population_and_social_statistics_Apr2021.pdf.
c. Economic and Social Commission for Asia and the Pacific, “Big data for environment and agriculture statistics”, Working Paper, Issue no. 13 (April 2021). Available at https://www.unescap.org/sites/default/d8files/knowledge-products/SD_Working_Paper_no13_Apr2021_Big_data_for_environment_and_agriculture_statistics.pdf.

[114] Chinese Academy of Sciences, “Big Earth data in support of the Sustainable Development Goals” (September 2020). Available at https://www.fmprc.gov.cn/mfa_eng/topics_665678/2030kcxfzyc/P020200927650108183958.pdf.

[115] See http://ggim.un.org/unwgic/nov20-ss-Measuring-Deqings-Progress-Towards-the-SDGs-with-Geo-Statistical-Information/.

[116] PARIS21, “Use of citizen-generated data for SDG reporting in the Philippines: a case study” (June 2020). Available at https://paris21.org/sites/default/files/inline-files/PSA-report-FINAL.pdf.

[117] For more information on country examples, see: Economic and Social Commission for Asia and the Pacific,” Big Data for the SDGs: Country examples in compiling SDG indicators using non-traditional data sources”, Working Paper, Issue no. 12 (January 2021). Available at https://www.unescap.org/sites/default/d8files/knowledge-products/SD_Working_Paper_no12_Jan2021_Big_data_for_SDG_indicators.pdf

8.3. VALIDATING AND IMPROVING OFFICIAL STATISTICS

164. It is possible to use external data sources for benchmarking or validation of survey results, or to use survey results for challenging alternate data sources, either at the micro- or macro-levels. In the validation process, the following actions should be considered:

- Assessing the origin and quality of the source, including trustworthiness and commercial or other interests of the parties exploiting them;
- Noting the concepts, definitions, classification and reference periods of the data and statistics being compared;
- Designing processes and modelling techniques that are sustainable and formalized (as ad hoc adjustments to the statistics would be difficult to defend);
- Educating users on proper use and interpretation of information (both the general public and more specific user groups).

165. Macro-integration is the process to integrate data from different sources on an aggregate level to enable a coherent analysis of the data. When there are two or even more independent data sources, inconsistencies will inevitably occur. Macro-integration can be divided into two stages. In the first stage the source data are adapted to comply with the correct definitions. The second part is called data reconciliation; this is the process in which the remaining discrepancies are resolved or at least reduced at the aggregated level. Data reconciliation is often called balancing in the literature. Data reconciliation, which may improve accuracy, can be divided into adjusting for major errors and adjusting for the remaining (sampling) noise. Large errors are often corrected manually by using subject matter knowledge. As correction methods for (large) errors are difficult to formalize, the coverage of them in the literature is limited. The ESSnet project on macro-integration discusses various methods of integrating data sources at the macro level.[118]

166. Data integration in various forms can be used to improve official statistics. The following list shows the opportunities that data integration has provided at Statistics New Zealand:

- The Integrated Data Infrastructure[119] brings together linked datasets from a range of government agencies (including the agency's own data collections). It is a large research database containing microdata about people and households and is continually expanding. The database has paved the way to answer complex research questions to improve outcomes for New Zealanders (for more details see the Integrated Data Infrastructure linking methodology [120] and the Integrated Data Infrastructure prototype spine's creation and coverage.[121])

[118] See https://ec.europa.eu/eurostat/cros/content/macro-integration_en.
[119] See https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/.
[120] Statistics New Zealand, "Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project", 4 July 2014. Available at https://www.stats.govt.nz/methods/linking-methodology-used-by-statistics-new-zealand-in-the-integrated-data-infrastructure-project.
[121] Andrew Black, "The IDI protype spine's creation and coverage", Statistics New Zealand Working Paper No 16–03 (July 2016). Available at

- The Longitudinal Business Database[122] is a large research database that holds de-identified microdata about businesses. Data comes from a range of Statistics New Zealand surveys and government agencies. The database complements the Integrated Data Infrastructure; researchers use it to evaluate policies and analyse business performance.
- Linked Employer-Employee Data [123] provides information on New Zealanders' interaction with the labour market and their sources of income. The longitudinal nature of the data allows analysis of income transitions, job tenure, multiple job holding, and self-employment. The data are created by linking a longitudinal series of the employer monthly schedule of the Inland Revenue to employer data from the Business Frame of Statistics New Zealand.
- Progress towards achieving the goals of the Statistics NZ's Census Transformation Programme[124], which is investigating alternative ways of running the country's future census, including the feasibility of using linked administrative data to replace census questions. In the 2018 census, administrative data were used to help compensate for the lower-than-expected participation.[125]
- Data integration has also paved the way for the development of new methods, such as new models. One good example is the production of population estimates applying Bayesian modelling to estimate, specifically, regional populations in New Zealand based on administrative data on birth and death registrations, tax and New Zealand international passenger movements.[126]
- Data integration has also assisted in the validation of income data from the Household Labour Force Survey with personal income available at Inland Revenue.

https://web.archive.org/web/20201025144316if_/http:/archive.stats.govt.nz/methods/research-papers/working-papers-original/idi-prototype-spine.aspx#gsc.tab=0.

[122] See https://www.stats.govt.nz/integrated-data/longitudinal-business-database/.

[123] See https://www.stats.govt.nz/methods/leed-annual-technical-notes.

[124] See https://www.stats.govt.nz/methods-and-standards/census-transformation-programme/.

[125] Statistics New Zealand, "2018 census: How we combined administrative data and census forms data to create the census dataset", 29 April 2018. Available at https://www.stats.govt.nz/methods/2018-census-how-we-combined-administrative-data-and-census-forms-data-to-create-the-census-dataset.

[126] John R. Bryant and Graham Patrick, "A Bayesian approach to population estimation with administrative data", *Journal of Official Statistics*, vol 31, No.3 (2015) pp. 475–487.

FINAL COMMENTS

167. Increasingly, new data sources are becoming available to official statistics organizations. These new data sources can be used to provide new official statistics, address new or unmet data needs, lower response burden, overcome the effects of declining response rates, and address quality, coverage and bias issues in surveys to meet national and international reporting commitments under the Sustainable Development Goals. The focus on data integration is increasing around the world.

168. This guide provides a starting point for developing or enhancing data integration expertise by reflecting on some of the issues faced and solutions found by some of the national and international official statistics organizations. Combined with the considerable discussions and resources gathered and available through the Data Integration Community of Practice (DI-CoP), the 2020 Data Integration Capacity Assessment Survey (DI-CAS), and the survey results along with the regional workshops on implementation of data integration in Asia and the Pacific, users of this guide can adapt, extend and contribute to work done by others in the region or globally for their own data integration endeavours. There are many examples, presentations and papers on both general techniques and projects related to specific statistical domains. To this end, users are invited to join DI-CoP,[127] and contribute to as well as explore and learn about data integration, with the aim of accelerating this practice in the region. In addition, as this is a live document, users are invited to share information and experiences related to data integration and point out any mistakes noticed in the guidelines.

[127] To join DI-CoP, click on https://stat-confluence.escap.un.org/pages/viewpage.action?spaceKey=DICP&title=Data+Integration+Community+of+Practice.

Chapter 9

ANNEX

Summary of Results of Data Integration Capacity Assessment Survey (DI-CAS)

Introduction

In June 2020, the Data Integration Community of Practice agreed to design and conduct the Data Integration Capacity Assessment Survey, based on a ECE survey*. The survey was conducted in September and October 2020.

Description of respondents

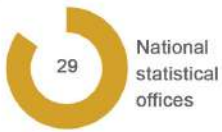
Countries responding to survey

Yes	49%
No	51%

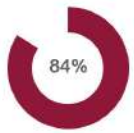
National statistical system of the country

Centralized	34%
Semi-centralized	24%
Decentralized	41%

Number of responding organizations: 31



Responding organizations that integrate data to produce statistics



Skills and capacity-building

Organizations having access to the required skills for DI

Yes	45%
No	55%

Organizations interested to receive training or invest in capacity development on DI

Law and privacy

Organizations having laws to access and use data from

	Government	Private sector
Yes	65%	52%
Partial	29%	23%
No	6%	25%

Due to privacy issues access to data by organization is:

Prohibited	32%
Partially restricted	35%

Strategies for Data Integration

Organizations having a DI strategy

Yes, in place	29%	Under consideration	26%
Under development	29%	No	16%

Organizations having:

specific units/functions to undertake DI projects	52%
specific units/functions to provide guidance or trainng on DI	40%
No specific unit/function	29%
DI governance committee	20%

Pre-integration practices

Collaboration in the preparation of administrative or statistical classifications used for data	77%
Cooperation agreements for transferring data	74%
Collaboration with data providers on the preparation of legal documents to establish/maintain use of data	58%
Having a long-term partnership with two or more institutions using the same data	52%

Barriers

Maintaining access to data sources	61%	Lack of supporting legislation	45%
Public acceptance and trust issues	52%	Not having access to data sources	45%
ICT issues	52%	Lack of relevant methodologies	42%

Identifiers

Responding countries having unified identity system in place

	Personal	Business	Address
Yes	72%	62%	45%
Partial	14%	17%	17%
No	14%	21%	38%

Data Integration practices

Prominent application of DI

Supplement surveys	80%	Social statistics	84%
Supplement traditional censuses	68%	Economic statistics	74%
Build sampling frames	61%	Environment statistics	35%
Data validation and/or imputation	61%		

Aspiration to use DI to produce or improve SDG indicators

Methods and tools

Widely used tools for linking

Excel	74%	SPSS	48%
SQL	68%	R	48%

Linkage method used

Deterministic	81%
Probabilistic	35%

Quality

Measurement of quality

As an integral part of quality reporting	65%
On an ad hoc basis	19%

Use of quality framework for DI activities

Yes	32%
No	42%

ECE survey: available at https://statswiki.unece.org/display/DI/2017+Data+Integration+Survey

DI: Data Integration

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