UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

ECE/CES/BUR/2023/OCT/3
22 September 2023

CONFERENCE OF EUROPEAN STATISTICIANS

Meeting of the 2023/2025 Bureau
Cardiff, UK, 9-10 October 2023

For discussion and recommendations

Item II (b) of the Provisional Agenda

IN-DEPTH REVIEW OF LINKING DATA ACROSS DOMAINS AND SOURCES – SCOPING PAPER

Prepared by Canada and Poland

In February 2023 the CES Bureau agreed to conduct an in-depth review on "Linking data across domains and sources" led by Canada and Poland. The review takes place in two steps: 1) a short paper including an outline on the work to be done is presented for discussion by the Bureau in October 2023; 2) a full paper, including information collected from countries and international organizations, presented for discussion by the Bureau in February 2024.

This note is the preliminary "scoping paper", introducing the focus of the review, the country context and the role of NSOs, lessons learned, existing issues and the next steps for the completion of the review before the final discussion by the Bureau, planned in February 2024.

The Bureau is invited to review the paper and provide comments on the continuation of the work.

I. INTRODUCTION

1. The aim of this note is to refine the topic and expected outcomes of an in-depth review on linking data across domains and sources, with this scoping paper presented for discussion at the meeting of the Bureau of the Conference of European Statisticians (CES) meeting on 9-10 October 2023 and the in-depth review to be carried out at the February 2024 meeting.

2. There is increasing demand from policymakers for data driven insights that address cross-cutting issues. In recent years, the issues at the top of global and national policy agendas, such as the COVID-19 pandemic, climate emergency and cost of living crises, have shown that the domains of society, economy, environment and health are connected and can no longer be treated in silos.

3. Linking data across domains and sources provides new opportunities for National statistical offices (NSOs) to develop a clearer picture of interrelated phenomena and address the need for disaggregated data both at specific population levels (e.g., diversity groups) and geographic levels, which provide granular information for the planning, administration and monitoring of policies.

4. Data linkage is not new. Countries have been linking administrative, survey and census data for many years. In recent years, data linkage has gone from record linkage of two data sources (e.g., census and survey data) to constructing a linkable environment of data files from

different administrative sources from across jurisdictions. Data linkage has also facilitated compilations of statistical information in different indicator frameworks. These new data linkages and statistical outputs pose new challenges in methods, toolkits and governance in data linkage, interoperability and stewardship.

5. The CES and the High Level Group for the Modernization of Official Statistics (HLG-MOS) have already completed a Guide to Data Integration for Official Statistics and an in-depth review in 2017 that documents the outcomes of the 2016 HLG-MOS Data Integration Project that involved twelve NSOs and Eurostat[1]. These materials offer a broad overview of the most common types of data integration that NSOs are presently conducting, and provide guidelines on the planning stages, data considerations, and methods and tools that pertain to each type. In addition, the CES and HLG-MOS have recently completed an in-depth review on data ethics that documents issues that are germane to data stewardship in the context of expanding data linkage. Work on a governance framework for data interoperability to improve future data linkages is currently undertaken under HLG-MOS.

6. Building on previous reviews, the present in-depth review will discuss **the readiness of NSOs for linking data** across domains and sources and highlight examples of **how NSOs can use data linkage to reposition themselves** from providers of data to producers of statistical indicators and multidimensional insights into complex social and economic phenomena. These examples are also intended to show how NSOs are well-positioned to be data stewards in a data ecosystem that has become increasingly complex.

7. **An important objective of this review is to raise awareness about the need for a systematic approach to data linkage.** The starting point for linking data should be a clear articulation of a question and follow the necessity and proportionality principles[2] so that the linked data are necessary for informed decision-making, while accounting for ethical considerations (Rancourt, 2019).[3] Data linkages that are implemented without a clear articulation of the information need and systematic approach increase the risk of producing data that are unsound and lack relevance.

8. Another objective of the review is to provide a platform to raise awareness of outcomes from previous CES projects relevant to linking data as well as bringing up future opportunities for international collaborations. This includes the 2016-2017 Data Integration Project, the 2018 Guidance on Data Integration for Measuring Migration, the 2023 In-depth Review of Data Ethics and the forthcoming Data Governance Framework for Interoperability.

II. FOCUS OF IN-DEPTH REVIEW

9. At the February 2023 meeting of the CES Bureau, the topic of "**Linking data across domains and sources"** was selected for an in-depth review. The CES Bureau members acknowledged that this topic is vast and the scope and expected outcomes of the review need to be refined. The CES Bureau has provided some preliminary comments on the potential areas of focus of the in-depth review:

[1] In 2017, the UNECE conducted a survey on the data integration practices of NSOs. The survey asked NSOs about their data integration activities across different domains and sources, their primary uses of integrated data, their methods, tools and quality frameworks for data integration, and the barriers to data integration.

[2] In 2019, Statistics Canada developed and implemented a Necessity and Proportionality Framework. More details about this framework can be found at Principles of Necessity and Proportionality (statcan.gc.ca) and a summary will be presented in the next section.

[3] Rancourt, E. (2019). The scientific approach as a transparency enabler throughout the data life-cycle. Statistical Journal of the IAOS, 35(4), 549-558.

(a) *A possibility is to focus on how* ***policy needs*** *can be met with by linking data, related to the paradigm shift of NSOs becoming producers of services and insights.*

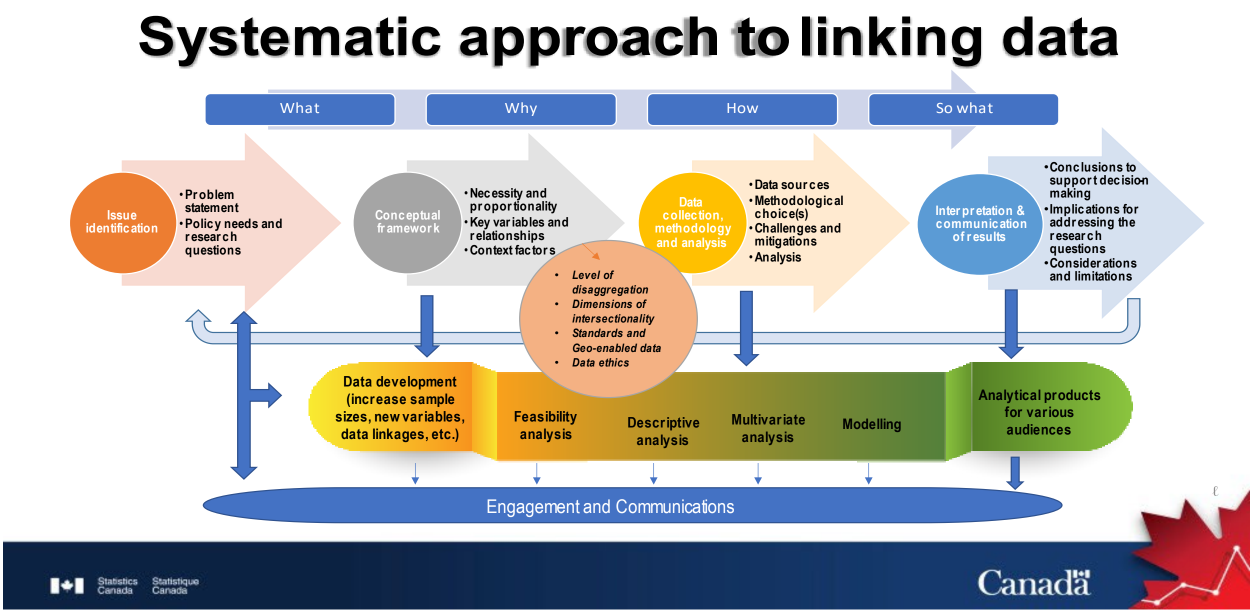
(b) *The starting point for data should be the aim for which data are needed, not the other way around. An underlying framework and a* ***systematic approach*** *are important to be able to link data so that it makes sense.*

(c) *A review could look at* ***examples of statistical outputs*** *produced with linked data: what kind of policy issues have been addressed successfully by linking which kind of data, what was the demand, output and impact?*

10. Accordingly, the main focus of this in-depth review is to highlight how NSOs have used data linkages in recent years to produce a clearer picture of interrelated phenomena, develop standard measures that work across frameworks and increase responsiveness to data needs when linking data across different sources and domains. This discussion also underscores the importance of having a systematic approach to linking data to meet policy, analytical and/or operational needs. An additional focus is to raise awareness about previous work that the CES and HLG-MOS have completed on linking data and related topics to avoid duplication, while directing attention to issues that need to be considered in the process of linking data.

11. The starting point for linking data should be guided by policy, analytical and/or operational questions or needs, informed by previous work and be purposeful throughout all steps of a systematic approach. This approach to linking data uses systematic checkpoints to ensure a sound implementation (Figure 1).

Figure 1
Systematic approach to linking data



Systematic approach to linking data

What

Why

How

So what

Issue identification

- Problem statement
- Policy needs and research questions

Conceptual framework

- Necessity and proportionality
- Key variables and relationships
- Context factors

Data collection, methodology and analysis

- Data sources
- Methodological choice(s)
- Challenges and mitigations
- Analysis

Interpretation & communication of results

- Conclusions to support decision making
- Implications for addressing the research questions
- Considerations and limitations

- *Level of disaggregation*
- *Dimensions of intersectionality*
- *Standards and Geo-enabled data*
- *Data ethics*

Data development (increase sample sizes, new variables, data linkages, etc.)

Feasibility analysis

Descriptive analysis

Multivariate analysis

Modelling

Analytical products for various audiences

Engagement and Communications

(a) The initial steps of a systematic approach to linking data are identifying an information need (i.e., what is the issue?) and establishing the necessity for linking data (i.e., why is linking data necessary for specific policy, analytical and/or operational needs?). These steps not only verify the relevance of the issue, but also provide the knowledge context so that the linked data will include the key variables and no more data will be linked than necessary.

(b) Subsequently, linking data involves several methodological steps, including the selection of the appropriate data sources (i.e., what data will be linked?), feasibility analysis and methodological choices (i.e., how will the data be linked?).

(c) The final step focuses on the communication of the expected benefits of the linked data along with a consideration of strengths and limitations. A feedback loop reinforces the connection between the outcomes of the linked data and new information needs.

(d) Continuous engagement with stakeholders is an essential part of a systematic approach to address information needs in a purposeful and rigorous way.

12. Overall, following a systematic approach to linking data facilitates the identification of information needs, knowledge gaps and relevant analytical questions, and promotes engagement with data providers, stakeholders and data users throughout all stages.

13. The approach and method to linking data also depends on the legislative, institutional and privacy context in each country. For example, in Canada, to help guide linking data across different domains and sources, **the principles of necessity and proportionality** were developed and implemented by Statistics Canada in 2019 to ensure that the linked data are necessary for informed decisions, while accounting for ethical considerations such as privacy, fairness and transparency.

(a) The Necessity and Proportionality Framework is an adaptation of the scientific approach to optimize privacy protection and the production of information.

(b) Necessity is the principle related to data needs, who requires the information, and the reasons why such information is needed.

(c) Proportionality is the extent of the effort needed to obtain the needed information in a manner that is coherent with the expected benefits of the project.

(d) Accounting for privacy intrusion, other ethical issues, and the data quality, the proportionality principle ensures that no more information than what is needed to produce the expected benefits is collected.

14. The next section will discuss examples of data linkages across domains and sources that meet different policy, analytical or operational needs.[4] These examples will also be used to identify challenges in different applications and approaches or methods that the NSOs have adopted to derive greater value in statistical insights.

III. COUNTRY CONTEXT AND THE ROLE OF NSOs

15. In most countries, the role of NSOs is essential given the level of technical expertise, IT infrastructure and data stewardship that is required to link the vast amounts of data that are available and needed to produce statistical information on cross-cutting policy issues. The centralization of data linkage within NSOs has many advantages for producing the best estimates possible (completeness and accuracy of data), responsivity to policy needs, data accessibility and comparability of indicators across frameworks, in addition to reducing response burden and survey costs.

16. Countries have been linking administrative, survey and census data for many years, but the scope of record linkage across domains and sources has greatly increased. The amount of data that is housed under one roof and the readiness for linking this data varies widely across

[4] In the current draft, three examples are discussed in Section III of the scoping paper.

countries. Countries with register-based systems are better prepared for data linkage across multiple domains and sources, whereas these data are more dispersed across jurisdictions and less interoperable in countries with decentralized systems.[5]

17. Register-based systems are based on administrative data from across domains (e.g., health, education, labour force) that are integrated into a statistical system. These systems streamline the process of data integration through having met key preconditions such as having a legal basis for data integration (e.g., a national Statistics Act), securing public acceptance of mass data linkage and implementing a unified identification system to integrate unit-level data from various sources.

18. In countries without a national registry system, data from different domains and sources are likely to have issues with interoperability as the data were not originally constructed to be used and examined together. The next two subsections discuss examples of how NSOs in countries without registry-based systems have overcome this limitation to bring together data from across jurisdictions.

A. Social and economic profile of persons who experienced opioid poisoning in Canada

19. For countries without national registries, there are more jurisdictional hurdles to access data and there is a technical need for a linkage environment to bring the data together. For example, in Canada, a country without a national registry system, data linkages across domains and sources face different jurisdictional challenges that involve data custodianship from multiple stakeholders. NSOs can take on a lead role to coordinate the data linkage while working with different data partners to put together a starting point for linkage across data from different domains and sources.

20. One example in Canada is the creation of an analytical file of persons who experienced opioid overdoses in British Columbia (BC). This file brings together longitudinal data from several provincial agencies (e.g., BC Medical Services Plan, BC Emergency Health Services and BC Corners Service), tax files from the Canada Revenue Agency (CRA), immigration data from Immigration, Refugees and Citizenship Canada (IRCC), and census and various sources of administrative data held at Statistics Canada. These disparate sources of data were integrated to provide a multidimensional view of the social and economic characteristics of persons who experienced opioid overdoses.

21. With these data managed by data custodians across jurisdictions, Statistics Canada took a lead role in the coordination, technical aspects and stewardship of the linkage. All the data that were agreed to be brought into the project were acquired by Statistics Canada and linked using a secure data linkage infrastructure to allow insightful analysis to be conducted in a confidential manner.[6]

B. The Integrated Metadata System in Poland

22. To improve data quality in a uniform manner for all units carrying out research, Statistics Poland has developed an Integrated Metadata System (IMS). The system consists of three lists

[5] This paragraph draws from the HLG-MOS Guide to Data Integration for Official Statistics. The UNECE has also published a report on Register-based statistics in the Nordic countries, which provides an overview of best practices with a focus on population and social statistics.
[6] Sanmartin, C., Garner, R., Carrière, G. et al. (2021). Statistics Canada British Columbia Opioid Overdose Analytical File: Technical Report. *Analytical Studies Methods and References*. Statistics Canada Catalogue no. 11-633-X – No. 031.

which includes a list of the population, a list of buildings and apartments, and a list of enterprises. For population list, seven data sources are used and one of the key processes is to collect and merge unique individual identifiers (PESEL numbers) from different administrative registers.

23. The IMS unlocks the huge potential of different administrative data and, on its basis, enables the creation of new analyses and observations of socioeconomic phenomena, including those previously unobservable. In addition to the main frame of persons, it also assumes the creation of separate thematic blocks (related to the frame of the population of persons), which contains substantive domain information concerning the entire population or a specific subpopulation and can be used to cover a variety of topics including migration, families, education, economic activity of the population, as well as housing and enterprises.

24. In the IMS, standardization of data used in official statistics are applied to all datasets. This example highlights the importance of investment and development of the processing in data linkage in order to improve the quality of the analytical findings and statistical outputs developed from the linked data.

C. Indicators framework – the Sustainable Development Goals of the United Nations

25. Linking data can also be used to support the development of indicator frameworks, which are designed to increase the accessibility of data for the surveillance of selected outcomes or policy objectives, such as the Sustainable Development Goals (SDGs) of the United Nations.

26. The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, has 17 Sustainable Development Goals (SDGs) which provides a shared blueprint for peace and prosperity for people and the planet, now and into the future. In 2017, a global indicator framework was adopted by the General Assembly on Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development. The goal of the SDG indicator framework is to monitor progress, inform policy and ensure accountability of all stakeholders. The 17 SDGs include 169 targets, which achievement is to be measures with 231 unique indicators. Action in one area will affect outcomes in others, and development must balance social, economic and environmental sustainability. Indicators are to be complemented at the regional and national levels.

27. Indicator frameworks compile summary statistics from across domains (e.g., health, education, income) to provide concise information on the relevance and connections between different indicators. However, the innovative use of existing and new data from multiple sources to inform indicators across frameworks may also result in the proliferation of indicators. In this context, the NSOs can take on the role of harmonizing or standardizing the same indicators so that these can be used consistently across frameworks. Having the NSOs take on the role of standardizing the indicators can also allow the indicators to be developed with the same subsets of the population, which is not always possible with unlinked data.

28. It is also important to note that it is not necessary to build new indicator frameworks with every new initiative. If a new framework is needed, it is important to communicate the purpose of the new framework and how the new framework relates to other existing ones so that users understand how various frameworks align, and where and how consistent indicators can be leveraged.

IV. LESSONS LEARNED AND OTHER CONSIDERATIONS

A. Lessons learned

29. In the in-depth report, the three examples discussed above (other examples can be suggested at the October CES Bureau meeting) will be elaborated on to illustrate the importance of the role of NSOs in linking data across domains and sources. Examples from Statistics Canada will also be used to show the necessity of following the systematic approach to linking data. Without a clear idea about what the linked data is being used for, it is not possible to identify the necessary data that is required for linking and the necessary concepts to be harmonized and standardized for the linkage environment.

30. In projects that are managed by a single country, it might be possible for NSOs to take on the lead role to apply a standardized linkage process as well as harmonizing the key measures and concepts – these roles have been taken up by Statistics Canada in the opioid project and Statistics Poland in the IMS.

31. However, in data linkages that are of larger scale, such as the SDG indicators, the international community needs to adopt widely accepted standards. For the SDG framework, a global SDG Indicators Data Platform was launched to provide a user-friendly interface to the Global SDG Indicators Database, access to the SDG Country Profiles and the SDG Analytics. The Analytics allows to review the availability of disaggregated data at the most elementary level. Annual refinements of indicators are included in the indicator framework.[7] These tools are important to ensure alignment of data, concepts, disaggregation categories and metadata.

B. Other considerations

32. Another role that NSOs often are entrusted with is to disseminate data for wider access so that more useful insights can be developed from different stakeholders including policy teams, academic researchers and private data users. There is often demand for new studies with the linked data in addition to what the NSOs have provided at the pilot phase. For the opioid project in Canada, in order to protect the sensitive information captured among the overdose population but also allow a wider use of this linked data to derive useful statistical insights, Statistics Canada is currently exploring ways to develop a synthetic database to allow increased access to the rich information from the linked data, while protecting confidentiality at the individual level.

V. RELATED ACTIVITIES UNDER CES AND HLG-MOS

33. CES and HLG-MOS have carried out several activities on the topic of linking data. This section provides an overview of these activities. This information provides context for the 2023/2024 in-depth review and avoids duplication of previous work.

34. CES conducted an in-depth review that described the outcomes of the **HLG-MOS 2016 Data Integration Project**. The aim of the project was to gain experience to "develop general recommendations and guidance for data integration and a related quality framework." The 2017 in-depth review on data integration focused on a broad-level discussion of the four most

[7] A SDMX-SDGs Working Group has been established to develop a global SDMX implementation framework for SDGs. The resulting global Data Structure Definitions (DSDs) and Metadata Structure Definitions (MSD) will be made available through the SDG Global Registry to be adopted for the collection and/or dissemination of official SDGs data and metadata to ensure timely and efficient collection, validation and dissemination of SDG indicators.

common types of data integration – (1) survey and administrative data, (2) new data sources (e.g., Big Data) and traditional data sources, (3) geospatial data and statistical information, and (4) integrating data for validating official statistics – with a brief description of country-level experiments with each type of linkage. The review also covered some of the main issues and challenges of data integration:

(a) Legal and institutional issues, such as data governance and the communication of policies on data confidentiality to assure public acceptance of data integration.

(b) Managerial issues, such as securing the human and IT resources needed for data integration as well as management of the risks of data integration.

(c) Methodological issues related to the harmonization of statistical concepts and definition of variables across data sources, data quality and procedures for dealing with missing data and linkage errors.

35. Importantly, the 2017 in-depth review emphasized that “using **standard processes** which are common for different types of data integration would greatly facilitate data integration.” The review provided a checklist list of what elements a standard process of data integration could include, such as identifying data needs, selecting data sources and analyzing the risk and benefits of data integration. This is an important starting point, but the notion of a “standard process” for linking data is under-developed. The current review will develop this idea under the systematic approach, adopting concepts from data ethics and applying the necessity and proportionality principles to linking data.

36. A task force of experts from NSOs developed the publication **Guidance on Data Integration for Measuring Migration**, which was endorsed by the CES in June 2018. Migration shapes societies. Its economic, social and demographic impacts are large and increasing. Policymakers, researchers and other stakeholders need data on migrants – how many there are, their rates of entry and exit, their characteristics and their integration into societies. These data need be comprehensive, accurate and frequently updated.

37. There is no single source that can provide such data on migration, but by combining several sources it might be possible to produce the information that users need. The publication provides an overview of the ways that data integration is used to produce migration statistics, based on a survey of migration data providers in over 50 countries. Thirteen case studies provide more detail on data integration in various national contexts. The publication proposes principles of best practices for integrating data to measure migration, presenting methods for combining administrative, statistical and other data sources for the production of migration statistics.

38. Data integration can also be used to improve migration statistics by developing longitudinal data sets. The data sources available to many NSOs have a great potential for producing longitudinal migration statistics. A UNECE task force of experts from NSOs and international organizations developed the 2020 publication **Guidance on the Use of Longitudinal Data for Migration Statistics**. A longitudinal approach where information is collected from the same individuals or households over time is particularly useful to understand migratory flows and the impact of migration on individuals, families, societies and economies.

39. The **2022-2023 HLG-MOS Data Governance Framework for Interoperability Project.** Data interoperability refers to the ability to efficiently exchange information and use it to be processed, integrated or disseminated. A Data Governance Framework for Interoperability (DAFI) aims to provide an overview of the governance elements that are required to be in control of information assets housed in NSOs and other organizations. The DAFI will describe the “core elements that are needed to establish and manage an interoperable

platform of data, metadata and systems." With the focus of the current review on linking data from different sources and domains, the recommendations from the DAFI will be highly relevant. The DAFI project is expected to conclude in December 2023.

40. The **2023 in-depth review of data ethics.** The issue of data ethics is not new, but the rapid expansion of data linkage reinforces the need for data ethics policies and raises new issues. For example, data linkages allow for the creation of large datasets with numerous variables describing individuals. This activity can be privacy-intrusive and done without the knowledge of the individuals on which data are gathered. This underscores the importance of necessity, proportionality and earning trust and social acceptance as well as enhanced data ethics processes and increased transparency throughout the data cycle.

41. This in-depth review provides insights on the ethical considerations that NSOs need to consider in linking data. The review includes examples of data ethics at several NSOs and Eurostat and guidelines on how to implement data ethics policies. A key message from this review is that the ethical considerations needed for linked data are broader than the ethical considerations for traditional data. In the traditional setting, NSOs have mainly focused on business ethics and data security. An enhanced view of data ethics is needed for data integration, which focuses on public acceptance in addition to issues of data confidentiality and security. With the proliferation of data across domains and sources, an ongoing concern is the potential for misuse of data. On surveys and censuses, the process of data collection is transparent as respondents know what information is collected on them and what it will be used for. This is less the case with administrative data and linkages between sources, and the public may not understand the value and scope of linked data or consent to its usage. This implies that an NSO needs to consider **what should be done** – not simply what can be done – to assure the social acceptance of linked data, which requires pro-active communication of the public benefits and a willingness to cancel projects where the future uses (or potential misuses) of the data are unknown.

42. Overall, each of the activities above provide useful and detailed information about specific topics that are important to consider in linking data across domains and sources. However, none of these activities focuses on taking a step back to see the big picture, which is the purpose of the present in-depth review. In particular, the present review builds on the need for standard processes in linking data.

VI. NEXT STEPS

A. CES Bureau discussion of the scoping paper

43. Previous work under CES and HLG-MOS has provided a broad overview of the key types of data integration and the issues that commonly arise with data integration as well as reviews on data interoperability and data ethics.

44. **CES Bureau members are invited to comment on the value-added of an in-depth review on linking data across domains and sources with regard to how NSOs can reposition themselves from data providers to producers of relevant statistical indicators and insights in response to the increasing need for multidimensional statistical information and indicator frameworks on complex social and economic phenomena.** As discussed in this scoping paper, the key issues for discussion at the CES Bureau meeting are:

- The need for a systematic approach to guide data linkage activities at NSOs.
- A discussion of the readiness of NSOs for large-scale data linkage.

- Examples of how NSOs can leverage their expertise to become providers of data insights and data stewards.

45. To move forward with the in-depth review, we would also like to receive input from the Bureau on the following questions:

- **Do the roles we have identified in the text from the NSOs resonate with your work? Are there other roles that NSOs can take on now or in the future?**
- **Approaches to data linkage depend on the contextual environment of different countries. Are there specific protocols or tools that your team has developed to facilitate data linkage across different domains and sources?**
- **Is there an international initiative or collaboration that NSOs can work with now or explore in the future to increase efficiency in compiling information from multiple sources?**

B. Collection of information from other countries

46. After the October 2023 Bureau meeting, it is planned to collect information from the countries that expressed willingness to contribute to the in-depth review and possibly from selected international organizations. Tentatively, the information to be collected could include issues such as:

- Country context and data requirement
- Role of NSOs/international organizations in taking up the lead
- Protocols/Tools developed
- Lessons learnt

47. The information collected from countries will be used to finalize the full in-depth review paper that will be discussed by the Bureau in February 2024.

* * * * *