United Nations ECE/CES/2024/5

Economic and Social Council

Distr.: General
11 April 2024

Original: English

Economic Commission for Europe

Conference of European Statisticians

Seventy-second plenary session
Geneva, 20 and 21 June 2024
Item 3 of the provisional agenda
Linking data across domains and sources

In-depth review of linking data across domains and sources

Prepared by Canada

Summary

This in-depth review was mandated by the Conference of European Statisticians (CES) Bureau and discusses the role of national statistical offices (NSOs) in linking data across domains and sources to meet information needs and produce statistical insights on multidimensional issues. The note provides an overview of a systematic approach to linking data and summarizes the experiences of NSOs in this area. The last section summarizes the discussion by the Bureau of CES at its meeting in February 2024.

The Conference is invited to endorse the outcomes of the in-depth review.

Please recycle

I. Executive summary

1. In February 2023, the Bureau of the Conference of European Statisticians (CES) agreed to conduct an in-depth review on "Linking data across domains and sources" led by Canada. This review raises questions for discussion about how national statistical offices (NSOs) can build on what they have already accomplished with data linkage, particularly with leveraging pre-existing data, new sources of data, and data science tools and techniques to improve efficiencies within national statistical systems (NSSs) and provide better information on policy issues and multidimensional problems. Issues at the top of global and national policy agendas, such as the COVID-19 pandemic, climate emergency, and energy and cost of living crises, illustrate the interconnectivity of the economy, society and the environment. Moreover, it is clear that impacts of these issues are not uniformly felt across societies and greater granularity of information is required to address inequalities. As a result, policymakers are taking a more holistic view of issues to address the interlinkages across domains. This is driving the demand for NSOs to provide statistical insights that address these cross-cutting policy issues and provide more granular information across multiple domains.

2. This review describes a systematic approach to linking data for the purpose of policy, analytical, or operational needs in NSOs and across departments. The review is based on case studies and a survey that focused on the readiness of NSOs to act as coordinators of data linkage activities in data ecosystems that are increasingly complex. The review also draws attention to previous CES work in this area.

3. Several themes emerged from the case studies and survey. Many data-related issues for NSOs have been solved by linking data; however, it is clear that data linkage can not only be a problem-driven (reactive) solution, but also an opportunity-driven (proactive) activity. Firstly, linked data are commonly used as problem-driven solutions to mitigate challenges related to declining response rates, missing data, and data quality. Secondly, linked data are used as opportunity-driven solutions to improve the efficiency of NSSs through reducing survey costs, response burden, and data redundancy in NSSs. Thirdly, linked data are a cost-effective means for generating data that is more frequent and responsive, disaggregated at subpopulation and geographic levels, and has the capacity to detect multidimensional social and economic phenomena that are invisible in single sources of data.

4. Challenges were identified in creating a data ecosystem that is efficient and serves the information needs of all stakeholders. At the root of these challenges was the lack of an organizational entity to coordinate the data assets in the NSS and develop an integrated system of linked data. One of the key recommendations of this in-depth review is the need for a user-gateway that enables access to as well as provides services to linked data. In absence of a coordination, there is a high risk of data fragmentation (lack of access), data duplication (redundancy and response burden), and data inconsistency in NSS as well as a lower capacity to observe interrelated phenomena.

5. There are also opportunities to harness sources of data not traditionally used by NSOs (e.g., satellite imagery and scanner data) to push the boundaries of what can be achieved with data linkage. However, these opportunities come with new challenges in acquiring these sources of data, which are often privately held, and require public acceptance for their usage. New technologies and tools (such as artificial intelligence (AI) and machine learning (ML)) can also be harnessed to better leverage linkages across domains, but risks associated with these technologies (ethical, technological, infrastructure) need to be addressed.

II. Introduction

6. The Bureau of the Conference of European Statisticians (CES) regularly reviews selected statistical areas in depth. The aim of the review is to improve coordination of statistical activities in the United Nations Economic Commission for Europe (UNECE) region, identify gaps or duplication of work, and address emerging issues. The reviews focus on strategic issues and highlight concerns of statistical offices of both a conceptual and coordinating nature.

7. The CES Bureau selected "Linking data across domains and sources" for an in-depth review at its February 2023 meeting. This review provides a summary of international statistical activities in the selected area, identifying opportunities and challenges, and makes recommendations on possible follow-up actions. The current review is led by Canada, with inputs from Poland and other countries who participated in a survey to inform the review.

8. Data linkage is not new and NSOs have been linking data for many years. However, the type and number of data sets being linked are evolving as is the domains across which data are being linked. This is being done to provide better insight into complex issues that require policy action across multiple domains. As noted below, several previous CES in-depth reviews have discussed **how** to link data, which ranges from technical issues to privacy and oversight challenges. This present in-depth review focuses on emerging issues on **why** there are needs for linking data and **where** there are challenges and opportunities for pushing data linkage into new frontiers. This in-depth review identified three areas for discussion at the CES meeting in February:

9. **First, data linkage can be used to create a new system for how NSOs produce official statistics and fill information needs.** There is a vast amount of data available within NSSs and NSOs have the opportunity to bring this data together to decrease their dependency on surveys for official statistics. Conventional data collection practices have started with surveys and later linked this data with other sources to fill gaps or improve data quality. This data collection process can be reversed to start with linking data, which can be described as a "link first, collect later" approach as it leverages pre-existing data first and later conducts surveys only as needed to supplement this data. This is consistent with the "once-only principle" in which the same data are collected once and reused across the NSS. For example, to reduce the time businesses spend responding to surveys, Statistics Canada uses administrative data that businesses and farms have already provided, such as tax returns and employee payroll records in surveys whenever possible. Other sources of data such as Remote Sensing and Traceability have also been assessed to substitute business surveys.

10. **Second, the complex information needs of policymakers are a driver of linking data across domains and sources.** The issues at the top of national and global policy agendas such as the climate emergency and the Sustainable Development Goals show that the domains of the economy, society, and environment are intertwined and issues within these domains should not be examined in isolation. Linked data are necessary for understanding the interrelated nature of these issues and the scope of their impact. For example, recent Canadian research based on linked data has shown that COVID-19 led to changes in work arrangements that have implications for public transit use and greenhouse gas emissions. Other data linkage projects have used disparate data sources (e.g., health services, coroner, income, and justice system) to provide detailed and nuanced insights into the opioid crisis in Canada that were invisible in single sources of data. Statistics Canada has also implemented the Real-time Local Business Conditions Index (RT-LBCI), which links Business Register data, with commercial data (Google Places, Yelp Fusion, and Zomato), and Road traffic data to create frequent (weekly or better) and geographically granular (city and neighbourhood) statistics to monitor business activities following the COVID-19 economic disruptions and recovery. Simply put, without linked data there is a high risk of information fragmentation or incomplete insights into complex issues.

11. **Third, the increase in the availability of new types of data and AI technologies provide new opportunities for linking and accessing data on scales that are more extensive than ever before.** New data sources are increasingly available – and necessary – for addressing multidimensional issues and providing data that is unavailable in conventional sources. A recent UNECE report has explored how linking data from conventional sources (e.g., surveys) to new sources of data (e.g., social media or mobile telephone data) can be used to develop better measures of migration and cross-border mobility. However, there are greater technical challenges with linking and analysing these new types of data. The High-Level Group for the Modernisation of Official Statistics (HLG-MOS) has already provided a framework for the quality of Big Data and guidelines for the establishment of partnerships in big data projects. There are many ways in which AI facilitates data collection, to categorize and make safe and effective predictions about data, and to enhance the value of data analysis. For example, Statistics Canada used Machine Learning (ML) to resolve issues of

interoperability (e.g., different ways of classifying data) that exist among similar types of administrative data (e.g., coroner data) from different jurisdictions. Another area that Statistics Canada has explored is the use of Privacy Preserving Linkage on sensitive data so that linkage and analysis can be done on the linked data in a federated model approach without the need of moving the data from one place to another. Further work is needed on how AI can be used to leverage Big Data for innovative uses and better information. There are also greater stewardship challenges in accessing this type of data and securing public acceptance of these large-scale linkages, which can be privacy intrusive.

12. Challenges remain for NSOs in data linkages, which range from technical issues needed for the feasibility of data linkage to stewardship issues needed for the oversight and public acceptance of data linkage. The overarching challenge that was identified in this in-depth review was the need for a user-gateway to provide access to linked data and services based on linked data. Given the vast amount of data providers and users within an NSS, there is a logistical necessity for a department or office that is responsible for creating an inventory of data assets in the NSS and coordinating data linkage activities based on FAIR (findability, accessibility, interoperability, and reusability) data principles.

13. The remainder of this in-depth review is organized as follows. Section III outlines the objectives of the review, which introduces a systematic approach for linking data. Section IV presents case studies and results of a survey of NSOs on their roles and experiences with linking data across domains and sources. Section V summarizes work on linking data that the CES Bureau has already completed. Section VI presents the issues and challenges that emerged from the review and Section VII concludes with recommendations for further discussion.

III. Scope of the statistical area covered

14. This in-depth review focuses on the role of NSOs as coordinators of data linkage activities in NSSs. In this review, data linkage is defined as record linkage or compilation of data from two or more sources, such as linkage of administrative sources from across government departments or ministries, linkage of survey and administrative data, or linkage of geospatial and survey data.[1] Linking data is not restricted to record linkages, though this is the most common type of linkage. Linking data can also involve compiling information from across domains, modelling, and the development of indicator frameworks. As brought up in the country examples of this review in the next section, linking data is a big topic which has many aspects, including domain-specific issues that could be embedded in the agenda of different thematic groups, and horizontal issues such as terminology, advocacy, social acceptability, techniques, and communication.

15. This review focuses on the reasons *why* NSOs are involved with linking data across domains and sources, and it does not discuss in great detail issues related to how data linkages are performed. The latter topic has been covered in prior CES work on the technological, methodological, managerial, legal and institutional challenges that are involved with linking data. This in-depth review has three objectives:

(a) First, to define and promote a systematic approach to linking data across domains and sources. This approach implements a recommendation from the previous CES in-depth review on data integration on the need for standard processes to facilitate data integration. The systematic approach formalizes a similar series of steps that can be used to guide data integration projects;

(b) Second, to document current and prospective roles of NSOs as coordinators of data linkages within NSSs and as providers of services based on linked data. The CES

[1] As noted in the UNECE comments on this review (see para. 10 in ECE/CES/BUR/2024/FEB/2/Add.1), terminology is an important aspect because in some cases the same terms may have different meanings in different countries. For example, in most countries, the term “administrative data” refers only to data that are maintained in the public sector for operational reasons while in some countries, the term refers to all data that are collected as part of an organization’s operations, including privately-held data.

requested that a survey on data linkage activities at NSOs be conducted as a feature of this in-depth review. The results from the survey highlight examples of how linked data have been used to serve policy, analytical or operational needs and the lessons learned from these experiences;

(c) Third, to raise awareness of relevant outcomes from recent CES work on linking data on a large scale such as an in-depth review of data integration (2017), an in-depth review of data ethics (2022), a report on data stewardship (2023), and a governance framework for data interoperability (in-progress). The present in-depth review briefly discusses some of these projects (Section V) as context for issues that were raised in the survey and are pertinent to the recommendations.

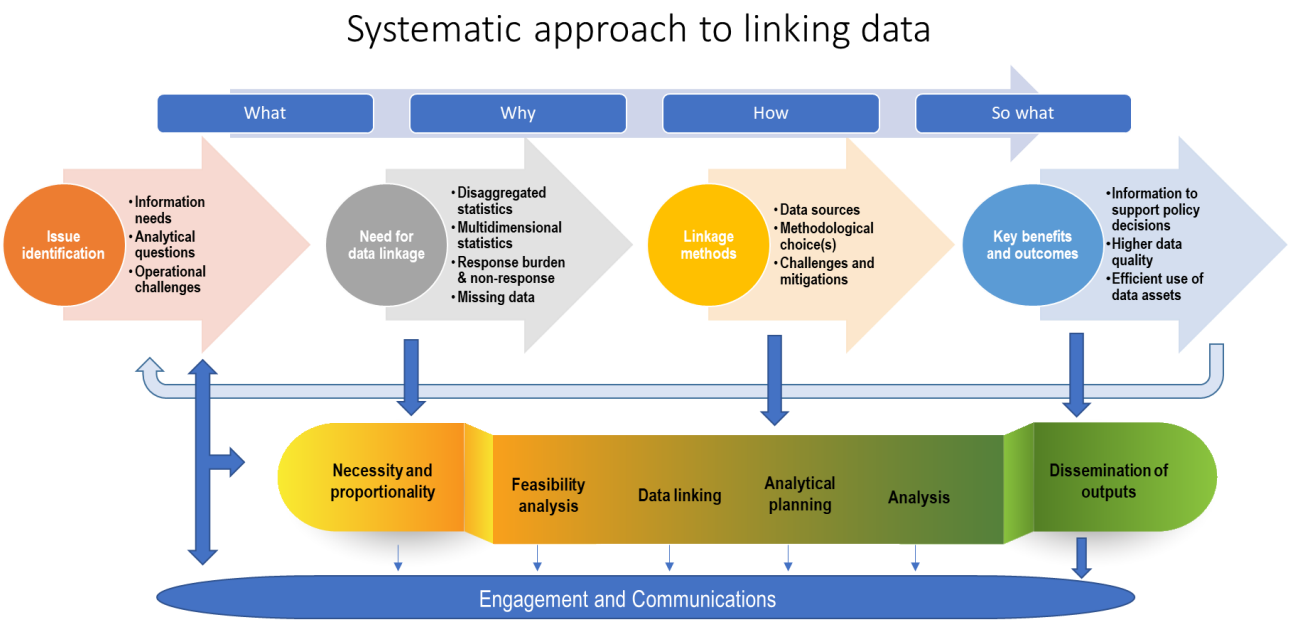
16. The review uses information from the case studies, survey, and previous CES work to show that NSOs are well positioned to provide a leadership role in the coordination of data linkage activities in NSSs. The review develops a set of recommendations of the issues and challenges that need to be addressed so that NSOs can function as user-gateways for linked data and services based on linked data.

Systematic approach

17. This in-depth review outlines the **need for a systematic approach to data linkage** to meet policy, analytical, or operational needs. The starting point for linking data should be a clear articulation of an information need, and follow the necessity and proportionality principles (discussed below) so that the linked data are necessary for informed decision-making, while accounting for ethical considerations (Rancourt, 2019). This is consistent with the recognition that social statistics "should be constructed as a system, knit together by a conceptual framework that reflects the best research knowledge and evidence on causal relationships, and the linkages between policies and outcomes" (Fellegi & Wolfson, 1999).

18. A systematic approach to linking data is guided by policy, analytical, and/or operational questions or needs, informed by consultation with stakeholders and evidence-based research, and purposeful throughout all steps. As visualized below, this approach to linking data uses a series of steps to ensure a sound implementation and user relevance.

Systematic approach to linking data



(a) The first step of a systematic approach to linking data is **issue identification** such as defining **what** information needs, analytical questions, or operational challenges (e.g., response burden, missing data) are to be addressed;

(b) The second step is the business case for **why** there is a need for data linkage, such as using data linkage as an efficient solution to the development of granular data and/or improvement of data quality. This step follows the principles of necessity and proportionality. *Necessity* is the principle related to information needs, who requires the information, and the reasons why such information is needed. **Proportionality** is the principle of how to obtain what data are needed (and no more than needed) in a manner that is coherent with the expected benefits of the project, while accounting for data ethics, confidentiality and transparency. This step is also informed by consultation with stakeholders to ensure the inclusion of essential variables and adequate coverage of subpopulation groups or geographic units;

(c) The third step focuses on **linkage methods** or **how** linked data will be used to address the data needs specified in the first and second steps. Different sources of data need to be evaluated so that the most useful sources are selected based on the variables and levels of disaggregation that are required. Key checkpoints of this step are identifying methodological challenges to the data linkage, approaches for mitigating these challenges, and inspecting the data to ensure sufficient sample size and data quality, while respecting data confidentiality. When data linkage is used to address an information need or analytical question, an analytical plan is needed to guide the analysis and align the project with stakeholder needs;

(d) The fourth step answers the **so what** question – i.e., this step communicates the **key benefits and outcomes** or value added of the data linkage, such as new data insights, higher data quality, reduction of response burden, or efficient use of data assets in the NSS. Effective communication of the benefits of data linkage is a core aspect of building and maintaining public acceptance of these projects. The outputs of data services based on linked data need to be clearly interpreted and distilled into key insights to inform decision-making and policy development. This includes a discussion of the strengths and limitations of the linked data.

19. A feedback loop reinforces the connection between the outcomes and new information needs that may emerge. From the outset, continuous engagement with stakeholders (both data providers and data users) is an essential part of a systematic approach to data linkage so that the outcomes will address information needs in a purposeful and rigorous way and also to foster a culture of data sharing in the NSS.

20. The approach and method to linking data is also dependent on the legislative, institutional and privacy context in each country. The next section discusses country examples of data linkages that meet different policy, analytical, or operational needs. These examples will also be used to identify challenges in different applications, approaches, or methods that NSOs have adopted to derive greater value in statistical insights.

IV. Country context and practices

21. Generally, countries that have a more centralized NSS are in a better position for statistical activities in linking data than are countries that have more decentralized systems. A major barrier to linking data from different sources is inconsistencies in the data and metadata due to a lack of standards or harmonization. An NSS cannot function as an ecosystem without **interoperability** or the logistical capacity to exchange and use data from across different sources in a coherent manner. Data linkage across sources may be further constrained in decentralized systems because of legal barriers to data sharing or a lack of commitment to horizonal initiatives as well as barriers related to coordination.

22. Countries that have register-based statistical systems have the highest degree of centralization and the Scandinavian countries have leveraged their register-based systems to become world leaders in linking data. 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, securing public acceptance of mass data linkage (social license), and implementing a unified identification

system to integrate unit-level data from various sources.[2] 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.

23. One disadvantage of centralized systems is that the NSO tends to be further removed from policy discussions in specific domains and may lack the subject matter expertise to produce user-relevant statistics and analytical products (Edmunds, 2005). Consistent with the systematic approach, ongoing consultative engagement is needed to set priorities and ensure that the outputs of data linkage are responsive to the information needs of the departments that compose the statistical system.

24. Data linkage is important in statistical production across countries. For example, in the United Kingdom of Great Britain and Northern Ireland, data linkage plays an important role in the statistical production across the government and in cooperation with the academia and other research bodies. The following subsections provide case studies of data linkage from Canada, Poland and the Russian Federation to illustrate the advantages of NSOs as information leaders in a complex data ecosystem. In addition, a survey of NSOs was conducted on this topic, and the results from the questionnaire are presented in the next section.

A. Canada

25. Statistics Canada has a coordinating role in linking administrative data from across sources outside the agency. This role includes identifying information needs, consultation with data providers and users, feasibility analysis, harmonizing data and ensuring data quality, and developing the IT infrastructure and institutional practices to protect the confidentiality of respondents in data linkages. The Social Data Linkage Environment (SDLE), the Business Linkable File Environment (B-LFE), and the Linkable Open Data Environment (LODE) are examples of large-scale linkage environments developed in Statistics Canada. A linkage environment is a secure data infrastructure and processing system that links deidentified unit-level records from multiple administrative data files on a custom basis – it is not a single, integrated data file as linkages of specific sources and variables are performed only as needed for approved projects.

26. In addition to general linkage environments, Statistics Canada has also developed linkable analytical platforms across sources and domains to address complex information needs. Some examples include the Canadian Employer-Employee Dynamics Database (CEEDD), the Canadian Census Health and Environment Cohorts (CanCHECs), and the Education and Labour Market Longitudinal Platform (ELMLP). CEEDD is discussed as an example of such data linkages providing advantages in a complex data ecosystem.

27. CEEDD provides matched data on employees and employers in the Canadian labour market to allow research on various topics including labour performance and mobility, industrial organization, and economic development and growth. CEEDD is not a single, integrated dataset – the data structure is a set of linkable files based on unique individual and business identifiers. The linkable files that compose CEEDD come from administrative data from Statistics Canada as well as from the departments of revenue, employment and social development, and immigration.

28. CEEDD is designed to protect the data in accordance with the confidentiality provisions of the Statistics Act, Statistics Canada's Policy on the Use of Administrative Data, and the Directive on Microdata Linkage and also adheres to the principles of necessity and proportionality defined above. All of the unit-level records are deidentified and stored on a secure server at Statistics Canada. The CEEDD data are not available to the public because of the confidential nature of the records. To facilitate accessibility, a partial version of CEEDD (called the Business-Employee Analytical Microdata) is available at Statistics Canada Research Data Centres, which are secure physical environments where accredited

[2] See the UNECE report *Register-based statistics in the Nordic Countries* for further information on this type of statistical system.

data users affiliated with the hosting organization, with pre-approved analytical projects, and having become Statistics Canada deemed employees can access the data.

29. CEEDD has created opportunities for policy analysts and researchers to fill information gaps that support policy goals. For example, Indigenous peoples and members of visible minorities[3] are designated groups in Employment Equity Act of Canada, but before CEEDD there was limited information to monitor their progress with regard to business ownership. This lack of information has been identified by stakeholders as a barrier to program development and business support for employment equity groups (Gueye, Lafrance-Cooke, & Oyarzun, 2022), and is a potential risk to the achievement of policy goals related to inclusion and diversity. CEEDD (augmented with a further linkage to census data) has been used to address information gaps on business ownership of Indigenous peoples and Black Canadians, providing a portrait of their presence in the Canadian economy, their sociodemographic characteristics, and the performance of their businesses over time in a comparative perspective.

30. CEEDD has also provided opportunities for Statistics Canada to produce multidimensional insights on the employment outcomes of Canadians in context of changes in the labour market. The CEEDD data have provided information on the impact of resource boom across the national economy in Canada, showing how migration and long-distance commuting have spread out the benefits and costs of this boom across a wide geographic region (Green et al., 2019). One advantage of the linked tax data is that it provides a more accurate measure of the wages of commuters and non-commuters and thus provides greater insights into how booms in resource-rich regions have spillover effects on wages in labour markets in other regions. Other Statistics Canada studies have used CEEDD to measure the "gig economy" in Canada and the characteristics of gig workers, providing insight into the increase in non-standard work arrangements (Jeon, Liu, & Ostrovsky, 2021), and how firm-level characteristics affect the economic integration of immigrants (Ci & Hou, 2017).

B. Poland

31. Statistics Poland has developed the Integrated Metadata System (IMS), an IT environment for generating a statistical register and leveraging data from across domains and sources. IMS enables the observation and analysis of phenomena that cannot be separately observed in any of the sources of data brought into the system. IMS consists of three lists which includes a list of the population, a list of buildings and apartments, and a list of enterprises. For the population list, seven data sources are used and one of the key processes is to collect and merge records using national personal identification numbers from across different administrative sources of data. In addition to the main frame of persons, IMS has separate thematic blocks (related to the frame of the population of persons), which contain substantive domain information concerning the entire population or a specific subpopulation and can be used to cover a variety of information needs. IMS can be used for annual comparisons or analysis of changing trends in demographic, social, and economic phenomena.

32. IMS includes:

(a) The Administrative Registers Processing System (SPRA) in which data obtained from official registers, public administration information systems, and non-public information systems are prepared, transformed, validated, corrected and integrated;

(b) The Variable Quality System (VQS) is used to facilitate advanced analysis of the quality of administrative variables and controls changes in register metadata;

[3] The term is used in this document as it is the official demographic category defined by the Employment Equity Act, which is currently under review (see Employment Equity Act Review Consultation). Statistics Canada has also conducted consultative engagements to identify the appropriate terminology for the visible minority concept (see Visible minority concept consultative engagement).

(c) The System of Statistical Operations (SOS) is used to create a population register by integrating selected administrative sources;

(d) The Domain Data Sets (DDS) are understood as unique thematic information blocks, the subjective scope of which is defined by SOS reports, and their subjective scope covers a specific subject matter that is the basis for observing social, economic and spatial phenomena.

33. Statistics Poland has used IMS to fill important and time-sensitive information gaps. During the COVID-19 pandemic, a problem arose with determining the resources of medical staff available in Poland. The statistical surveys carried out were based on the reports of medical institutions. The data provided by them was aggregated and meant that if a doctor or nurse worked in several places, they appeared in the resulting data many times. Another phenomenon falsifying the data was the fact that part of the medical staff living close to the state border often took up work in a neighbouring country. In addition, some doctors, and more often nurses with a license to practice, worked in places other than medical facilities. With the integration of data from many sources (e.g., the register of doctors, the register of nurses, the social security system, the register of medical facilities), it was possible to make very precise and almost real-time calculations. This made it possible to determine how many doctors and nurses live in Poland, whether they have a license to practice, whether they work with patients, and in what type of facilities. It was also possible to present the demographic characteristics of these people and their territorial distribution.

C. Russian Federation

34. The Federal State Statistics Service of the Russian Federation (Rosstat) has developed the State Information System called the "Digital analytical platform for statistical data provision" (the Platform). The Platform implements the principles of (a) transparency in collecting and processing information, (b) use of a unified methodology for collecting and processing statistical data, (c) reliability and consistency of system information, and (d) single-point data provision and their multiple use. The Platform's work involves interaction with the Rosstat information and computing system as well as external information systems. The ability to download, process and analyse data (including official statistical information, administrative and primary statistical data) from various sources is implemented in the Platform subsystems. The Platform has been used for:

(a) **Calculation and analysis of indicators for achieving the national priority projects of the Russian Federation.** The purpose is to assess project effectiveness as well as the national development goals of the Russian Federation. The data sets used for calculating the indicators are configured automatically from external information systems. The system provides algorithms for calculating the indicators based on approved methods. Correlation analysis is implemented to assess the connection strength between the indicator components and their impact on the indicator itself;

(b) **The digital methodologies subsystem** automates the processes of formation, coordination and approval of methods for calculating indicators and converting them into structured electronic documents. The purpose is to unify and systematize the requirements for calculating national and federal project indicators, ensuring use of generalized approaches aimed at reproducibility and traceability of data. The subsystem is designed to (i) streamline the data harmonization process; (ii) provide a single point for collecting, calculating and reporting indicator values; (iii) evaluate progress toward achieving the indicator values, accounting for factors that influence them; and (iv) synchronize decision-making processes within project management and statistical production;

(c) **Technological module for rapid assessment of population income on 10 per cent (decile) group basis**. The module aims at improving formation and release of statistical information and developing measures to reduce poverty. The system accumulates data from various sources including survey microdata, output tables of official statistical observation forms, laws and regulations data, administrative data and supplemental information;

(d) **Social care microsimulation system**. This microsimulation predicts changes in key indicators under various scenarios of changes in population income and the poverty line. The system also determines the most effective measures of social support and budget expenditures to obtain optimal results. The microsimulation system is based on sample observation of population income data. It allows for the creation of various scenarios for change in individual-level income based on the subsistence level and poverty line changes. The purposely designed customer interface allows a user to simulate earnings from social insurance benefits and other income components such as employment or property income. The simulation result is a set of interactive analytical reports consisting of calculated indicators (change in level of poverty, budget expenditures, and per capita incomes of people). There is also an option for simulation and forecasting based on statistical data in various fields of statistics.

D. Survey results

35. As part of this in-depth review, a survey was distributed to NSOs who previously expressed interest in contributing information on their experiences with linking data across domains and sources. The survey included questions on (a) type of NSS, (b) current roles in data linkage, (c) prospective roles in data linkage, (d) examples of using data linkage for information needs, (e) protocols, tools, and infrastructure to facilitate data linkage, (f) lessons learned in data linkage, and (g) opportunities for international collaboration. Responses were received from Canada, Estonia, Hungary, Italy, Latvia, Mexico, and the Netherlands. The responses are summarized below and presented in detail in document ECE/CES/2024/INF.1.

36. **Type of NSS (country context).** Country context is an important dimension of the readiness of NSOs for linking data across domains and sources. The survey asked respondents to provide information about type of statistical system that their country has, including details on the degree of centralization in their systems and the role of their NSO in the coordination and production official statistics. The centralization of official statistics is common in the NSSs of the countries surveyed. Most NSOs who participated in the survey have legal access to administrative data from across the NSS and coordinate statistical activities. However, even in centralized systems, the data and responsibilities for official statistics in some domains are dispersed across departments and levels of government.

37. **Current roles of NSOs**. The questionnaire asked respondents to describe the current roles and responsibilities of their NSOs in linking administrative data from different sources. The NSOs surveyed generally have a coordinating role in linking data from across departments, and the use of these linked data are restricted to statistical purposes and regulated by laws that ensure the confidentiality of the data. Most NSOs have developed the IT infrastructure needed for data linkage and adhere to Statistics Acts to ensure both the confidentiality of the data and that the data are used for only statistical purposes and not misused.

38. **Prospective roles of NSO.** The questionnaire also asked respondents about whether there are additional roles that their NSO could take on now or in the future that would allow them to reposition themselves from providers of data to producers of relevant statistical indicators and multidimensional insights. It was reported that NSOs can take on an enhanced role in coordinating data linkage activities in the NSS. An enhanced role would support the efficient use of resources (data assets, human capital, and IT infrastructure) and exchange of information in the statistical system. The prevailing theme in the responses was that NSOs are well positioned to be user-gateways for access to linked data. The NSO can position itself as the “national analytical and competence centre responsible for administrative data linkage and providing access to the linked data” within a secure environment (Latvia).

39. **Data linkage for information needs.** The questionnaire asked respondents to provide examples of information needs in their country that have been solved by linking data. Most NSOs who participated in the survey have used data linkage for problem-driven needs related to operational challenges, such as reducing respondent burden, decreasing survey costs, increasing the precision of statistics, and dealing with coverage errors and declining response rates. Opportunity-driven linkages included the production of more timely statistics and

insights on phenomena that cannot be observed with a single source of data. NSOs reported that "data linkage is the most valuable service we have – it is faster and more flexible that official statistics" (Estonia) and it "provides multiple solutions to the problem of publishing instant and precise demographic statistics" (Hungary). "Probably the most important example of data linking is the System of Social Databases (SSB)" (the Netherlands). The SSB is opportunity-driven in that census statistics are based on data already available in the system.

40. **Facilitating data linkage**. The survey asked respondents whether there are specific protocols, tools, or infrastructure (e.g., in legal, IT, or data processing) that their NSO has developed to facilitate data linkage across domains and sources. The NSOs surveyed have (a) taken steps to facilitate the interoperability of administrative data in their NSSs, (b) developed the IT infrastructure needed to link and share the data, and (c) developed processes for linking administrative data to reduce redundancy and other inefficiencies in their statistical systems. This includes the development of linkage environments (secure data infrastructure and processing systems for linking de-identified records) and open-source software and ecosystem solutions for data sharing. The IT infrastructure at NSOs have been also implemented to enable the once-only principle, which aims to collect the same data once-only to avoid duplication and reduce response burden.

41. **Lessons learned.** The respondents were asked to provide details on lessons learned from their experiences with data linkage for information needs and interoperability. Two themes that stood out were the need for unique identifiers for data linkages and streamlined data sharing agreements across the NSS.

(a) **Value of unique identifiers.** A common ID code or single registry across administrative data is of general benefit for the production of official statistics and fosters data linkages that yield innovative products that meet the specific requirements of decision-making processes and also provide more detailed and frequent data that is available in surveys. The lack of unique identifiers is a challenge, and probabilistic linkages are a potential solution, but further work needs to be done to test these methods;

(b) **Need for enhanced data sharing protocols**. NSOs remarked that "Obtaining administrative sources is still too slow and difficult" (Italy) and that there is a "risk of losing the data flow" (Mexico) because of the need to establish data sharing agreements with third parties. Current legislative frameworks are insufficient for accessing data in a timely fashion and for accessing data from the private sector, which is needed "for better and more diverse data" (Estonia). Enhanced data sharing agreements are needed to formalize relationships between the NSO and other departments and to social acceptance of the linkage of administrative data and use of data from the private sector. To mitigate privacy concerns to link microdata from different sources, Privacy Preserving Record Linkage (PPRL) has the potential to allow analytics to take place on sensitive datasets without the need to move the data out of the custodian institutions (Canada).

42. **International collaboration opportunities.** Respondents were asked if there is 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. The respondents identified several ongoing international activities on the technical aspects and challenges of data linkage that are important initiatives. The HLG-MOS Applying Data Science and Modern Methods Group and Supporting Standards Group were singled out as important ongoing international collaborations.

V. Related work under the Conference of European Statisticians and High-Level Group for the Modernisation of Official Statistics

43. CES and HLG-MOS have carried out several activities on the topic of linking data. This section provides a brief overview of these activities. The information in this section provides context for the discussion of the issues raised in the present in-depth review and avoids duplication of previous work on this topic.

44. CES conducted an in-depth review that described the experiences gained from 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 in-depth review on data integration was presented at the CES meeting in February 2017 and focused on a high-level description of the most common types of data integration and country-level experiments with each type. The review also covered challenges of data integration, such as:

(a) **Legal and institutional issues** – i.e., legislation on confidentiality, communication of activities to assure public acceptance, collaboration with data providers, and oversight of data integration projects;

(b) **Managerial issues** – i.e., the human resources and IT infrastructure needed for data integration, and the organizational protocols needed to mitigate the risks that are inherent in data integration;

(c) **Methodological issues –** i.e., problems to be overcome such as a lack of unified identifiers, differences in the concepts and classifications used to define and organize the data, and missing data and coverage errors.

45. Importantly, the in-depth review recommended 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. The systematic approach outlined in Section III of the present review builds on this recommendation.

46. A task force of experts from NSOs developed the publication *Guidance on Data Integration for Measuring Migration*. Policymakers, researchers, and other stakeholders need data on migrants – how many there are, their rates of entry and exit, their characteristics, and integration into societies. These data need to be comprehensive, accurate and frequently updated. There is no single source that can provide such data on migration, but by combining several sources together 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.

47. The Data Governance Framework for Interoperability (DAFI) was developed through the HLG-MOS Data Governance Framework for Interoperability Project carried out in 2022-2023. Data interoperability refers to capacity to exchange and make use of the information with minimal or no prior communication. It is also the basis for continuous flows of data between sources and transforming siloed data assets "into a connected network of harmonized data and metadata sets.". DAFI describes 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, the recommendations from DAFI are highly relevant.

48. In 2023, the CES Bureau conducted an in-depth review of data ethics. 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. 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 different sources of data. 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 public acceptance of linked data, which requires communication of the public benefits and a willingness to cancel projects where the future uses (or potential misuses) of the data are unknown.

VI. Issues and challenges

49. The NSOs that participated in this review all reported having a coordinating role in the production of official statistics and legal access to administrative data from other government departments. However, NSSs are complex data ecosystems with numerous data providers and users, which poses challenges for data access and efficiencies across the NSS. Below are summarized themes that emerged from the case studies and survey of NSOs in their experience in linking data across domains and sources.

A. Current and future roles for national statistical offices

50. **Current data linkage activities at NSOs are aimed at efficient use of pre-existing data in the NSS to produce up-to-date and accurate information.** Increasing costs of surveys and decreasing response rates are a challenge for many NSOs, which is occurring alongside increasing demand for rapid response, disaggregated and multidimensional data. The NSOs in this review described practical reasons for data linkage, which corroborate the benefits of data integration outlined in the previous CES in-depth review on this topic:

(a) Common problem-driven data linkages are based on the once-only principle in which administrative data are integrated into surveys to replace variables (or even replace surveys entirely, such as is the case with register-based censuses) to solve the problem of data duplication across the NSS, which decreases survey costs and response burden. Linkages of administrative data have also been used to generate sampling frames for surveys and to resolve data quality issues associated with increasing non-response rates and coverage errors;

(b) Opportunity-driven linkages are used to:

(i) Harness the continuous flows of administrative data as a resource to produce more frequent estimates, monitor trends more closely, and respond to crises such as the COVID-19 pandemic;

(ii) Leverage the large size of administrative data to improve the coverage of small and hard-to-reach populations, fulfil data needs for disaggregated statistics, and observe phenomena that are invisible in single sources of data.

51. **National statistical offices are well-suited to operate as user-gateways for access to linked data and services based on linked data.** There was agreement among survey respondents that NSOs should be appointed a well-defined coordinating role because of the complexities of data linkage in NSSs that consist of numerous parties and increasing amounts of administrative data. It has become impractical for data-sharing agreements to be established on a bilateral basis and the human and IT resources needed to properly link data are too costly to reproduce multiple times across the NSS. NSOs have the necessary competencies to develop the infrastructure, methodologies, processes, and protocols needed to link data, while assuring FAIR data principles and data ethics.

B. Challenges for national statistical offices

52. **A lack of clearly specified roles and a streamlined process for data sharing is an impediment to linking data across domains and sources.** Having a centralized statistical system and legislation that permits NSOs to access administrative data from across NSS are insufficient conditions for NSOs to become user-gateways for linked data and services based on linked data.

(a) While NSOs who responded to the survey for this review generally had legal access to administrative data, there was less control over how these data are collected and the data structure, which can decrease interoperability and data quality;

(b) The lack of unique identifiers was a frequently mentioned challenge to data linkage across external sources, and the use of probabilistic linkage was discussed as a stop-gap approach that can deteriorate the quality of record linkage;

(c) A common process for how data are transferred, transformed, and linked is necessary to avoid repeating this work for every database in the system;

(d) Common standards for data collection, data structure, methodologies, identifiers, vocabularies, and definitions are not always present, but are essential for interoperability and decreasing the time and resources needed for data linkage;

(e) The process for obtaining administrative data from other departments was often slow and cumbersome.

53. The risks associated with deficient data-sharing protocols are increases in implementation time and losing data flow. Common data standards are needed to boost the capacity of the data to be used for future as well as current operational needs, and for purposes other than for which these data were initially collected. Improved standards and harmonization of data (including metadata) across domains and sources facilitate data linkage as well as support users who may be experts in their own domains, but not in others.

54. **There is demand for sharing linked data with businesses and non-governmental organizations, but this involves additional challenges.** In particular, sharing data with parties outside the public sector is a new frontier. Businesses are increasingly requesting data and data services from NSOs, but sharing data with businesses raises new issues related to public acceptance, data ethics and privacy that need to be resolved. There are many instances where sharing data with the private sector or non-governmental organizations is prohibited. Considering this asymmetry, the NSOs need to find a right balance to data linkage and also find ways to improve data sharing as part of a larger process of optimizing and modernizing statistical production.

55. **There is growing awareness that Big Data and privately-held data are needed to address information needs.** There is a need for NSOs to access data generated by businesses and ICT technologies (e.g., web scraping) to increase the capacity of official statistics to respond to economic issues in the private sector and monitor social issues not covered in administrative or survey data. CES completed an in-depth review on collaboration with private sector data providers that outlines the challenges and lessons NSOs have learned so far on this topic.

56. **The use of existing and new data from multiple sources to develop indicators may result in a proliferation of indicators.** NSOs can take on the role to standardize or harmonize indicators to ensure their consistency and comparability across indicator frameworks. Having NSOs take on this role also allows the indicators to be developed with the same subsets of the population, which is not always possible with unlinked data.

57. As noted in the challenges highlighted above, in addition to cultural changes that needed to be adopted from the NSOs, responses from data holders of administrative and other data sources are also necessary and important to allow NSOs to reposition themselves from data providers to producers of relevant statistical indicators and insights.

VII. Conclusions and recommendations

58. This in-depth review provided an overview of how NSOs have used data linkage to solve operational problems (e.g., declining response rates) and fill information needs, while dealing with the challenges of data sharing. The review also discussed a forward-looking perspective on how NSOs can reposition themselves from providers of official statistics to assuming additional roles as user-gateways for linked data and providers of insights on multidimensional phenomena. The review was based on case studies, a survey of NSOs, and research of previous CES work on data integration and related topics.

59. The following recommendations are made:

(a) **A systematic approach is needed for data linkage projects.** Data linkage is a key method that can be used to address complex information needs but requires a systematic approach to ensure sound implementation and policy relevance. This approach should be (i) guided by policy, analytical and/or operational questions or needs, (ii) informed by previous work, (iii) purposeful throughout all steps, which can be either problem-driven (reactive) or

opportunity-driven (proactive), and (iv) involve ongoing consultative engagement with stakeholders to support a cooperative environment on data sharing and standards and the user relevance of the linked data;

(b) At the core of the systematic approach, the existing including alternative data sources should be evaluated first so that the most useful sources are selected based on the variables and levels of the required disaggregation. **A change in the mindset for how data are collected is also needed.** NSOs should explore options of leveraging existing data sources first and then assess the fitness for use of other supplementary sources of data from different sources or domains;

(c) **There is a need for NSOs to have a formalized coordinating role in data linkage activities.** Legal access to administrative data from across the NSS does not eliminate barriers to accessibility of data from across sources, which has implications for the services that NSOs can provide and fulfilling information needs across the NSS. There are several advantages to providing NSOs with the capacity to operate as user-gateways for linked data. This gateway can be the hub for access to linked data across the NSS and avoids repetition of linkages. This also ensures the once-only principle so that data assets in the NSS are used efficiently, and supports the consistency of linkages and decreases the proliferation of dissimilar indicator frameworks;

(d) **The development of a roadmap should be considered to help NSOs looking to link data across domains to provide better information to policymakers as they seek to address multidimensional issues in society.** This roadmap should provide 1) guidance related to the governance that should be in place to undertake expanded data linkage across domains; 2) examples of the types of data across domains not traditionally used and that could be useful and linked to better inform policy development, and; 3) concrete country examples of multidimensional data linkages in policy areas that have been done to illustrate linking across domains;

(e) **Previous CES work on data integration and complementary issues can be leveraged to inform the development of such a roadmap.** The results from survey conducted for the present in-depth review indicated the technical challenges of interoperability and data stewardship challenges of data sharing agreements and securing public acceptance of linked data. The respondents of the survey conducted for the present in-depth review recommended the HLG-MOS Applying Data Science and Modern Methods Group and the Supporting Standards Group as important international collaborations on the technical aspects of linking data across domains and sources. Previous CES work on interoperability, data ethics and data stewardship also provide guidance for streamlining data sharing in the NSS and securing public acceptance of data integration projects. Rather than creating a new group to undertaking the development of a roadmap, existing groups could possibly be tasked to do this work.

VIII. Discussion by the Bureau of the Conference of European Statisticians

60. The Bureau conducted an in-depth review of linking data across domains and sources in February 2024 based on a paper prepared by Canada and comments by UNECE. The following comments were made:

(a) The paper gives a very good overview of the issues related to linking data focusing on strategic and managerial issues: the need for data linking and the related challenges and opportunities. The proposal to develop a road map would be a constructive way forward in international work on this topic;

(b) Linking data plays a key role in the repositioning of NSOs from data providers to producers of relevant statistical indicators and insights in response to the increasing need for multidimensional statistical information. This is an existential issue and requires a cultural change from NSOs but also from the holders of administrative and other data sources. Data linking is not new but the type and number of data sets that can be linked makes a

transformative change possible. It will remain a strategically important issue for years to come;

(c) A systematic approach is needed to change the mindset instead of only reacting to requests. NSOs need to have a formalized role in data linking and could become user gateways coordinating data linkages within their national statistical systems;

(d) The level of social acceptability of data linking in countries is different: in some countries it is well accepted while in some others it raises concerns. In countries where there is no system of unique identifiers or no population register, establishing data linkages poses significant challenges. This further emphasizes the importance of communication, transparency and partnerships;

(e) NSOs may be entrusted with linking data from different sources but not allowed to hand over the linked data to other institutions for legal or public acceptance reasons. This asymmetry may be difficult to accept to other parts of the government. The institutional setup between NSOs and the rest of the administration should have the right balance;

(f) Metadata, terminology and classifications need to be better harmonized to facilitate linking data. For example, the term “administrative data” has different meanings in different countries;

(g) Data linking is related to data sharing. They should be addressed at the same level of importance, driven by user demands. The horizontal and demand-specific issues need to be looked at the same time;

(h) Data linking allows to produce integrated data across domains. This increases the need to support users, who may be experts in their own domains, but not in others;

(i) Linking data is a big topic with many aspects, including horizontal issues such as terminology, advocacy, social acceptability, techniques and communication, and domain-specific issues that could be embedded in the agenda of different thematic groups;

(j) Rather than creating a new group, the discussion of horizontal issues and the development of a roadmap should be considered by the High-Level Group on the Modernisation of Statistics (HLG-MOS) and its groups on Applying Data Science and Modern Methods, and on Supporting Standards;

(k) A short questionnaire could be prepared identifying how different countries link data. On this basis case studies and examples could be collected;

(l) In the United Kingdom, data linkage plays an important role in the statistical production across the government and in cooperation with the academia and research. The ONS data linkage team would like to be involved in future work on this topic;

(m) CIS-Stat is willing to participate in developing a road map.

61. The following conclusions were reached by the Bureau:

(a) The HLG-MOS will consider including horizontal issues related to linking data in its work programme and in the agenda of its groups – namely those on Applying Data Science and Modern Methods and on Supporting Standards – as far as possible and whenever those issues are related to the mandates of the respective groups;

(b) The Bureau invites HLG-MOS to develop a road map on linking data, based on the outcomes of the in-depth review;

(c) Issues related to linking data across sources and domains should be mainstreamed in the programme of work of subject-matter groups working under CES and included in the agenda of expert meetings whenever relevant. This could entail collecting examples from countries to be shared and disseminated;

(d) The Bureau will follow up on the progress on this topic in the coming years.

References

Ci, W., & Hou, F. (2017). Immigrants' initial firm allocation and earnings growth. *Canadian Studies in Population*, *44*(1–2), 42–58.

Conference of European Statisticians Task Force. (2023). Data stewardship and the role of national statistical offices in the new data ecosystem.

Edmunds, R. (2005). Models of statistical systems. Paris: OECD.

Fellegi, I., & Wolfson, M. (1999). Towards systems of social statistics – Some principles and their application in Statistics Canada. *Journal of Official Statistics*, *15*(3), 373–393.

Green, D., Morissette, R., Sand, B. M., & Snoddy, I. (2019). Economy-wide spillovers from booms: Long-distance commuting and the spread of wage effects. *Journal of Labor Economics*, *37*(S2), S643–S687.

Gueye, B., Lafrance-Cooke, A., & Oyarzun, J. (2022). Identifying Indigenous business owners and Indigenous-owned businesses. *Analytical Studies: Methods and References*, No. 045 Statistics Canada, Catalogue no. 11-633-X.

Jeon, S-H., Liu, H., & Ostrovsky, Y. (2021). Measuring the gig economy in Canada using administrative data. *Canadian Journal of Economics*, *54*(4), 1638–1666.

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