United Nations ECE/CES/2024/INF.1

Economic and Social Council

Distr.: General
1 May 2024

English only

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 – detailed results of the survey

Prepared by Canada

Summary

As part of the in-depth review of linking data across domains and sources, 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.

This document, to be considered in connection with the in-depth review paper (doc. ECE/CES/2024/5), presents in detail the responses received from Canada, Estonia, Hungary, Italy, Latvia, Mexico, and the Netherlands.

1. Survey responses were received from Canada, Estonia, Hungary, Italy, Latvia, Mexico and the Netherlands. Below are summarized the key issues and themes that were identified from a content analysis of the survey responses.

2. **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 the 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. As highlighted below, the centralization of official statistics is common in the NSSs of the countries surveyed.

- **Canada** has a national statistical office, Statistics Canada**.** The agency has a legislated mandate under the *Statistics Act* to collect data on a wide range of social and economic topics and also to access documents or records maintained in any department, municipal office, corporation, business, or organization. The duties of Statistics Canada include collaborating with departments in the production of official statistics, promoting the avoidance of duplication of statistics across departments, and developing and coordinating plans for the integration of these statistics. A large portion of the administrative data used at Statistics Canada come from provincial and territorial departments who are responsible for statistics in certain domains such as education, health and the courts (which are sub-national systems) and other federal departments for data in domains such as immigration and income.

- **Estonia** has a national statistical office (Statistics Estonia) and a national registry-based system.

- **Hungary** has a Central Statistical Office (KSH) that is responsible for official statistics, but there is an Official Statistical Service (HSSz) which is a formal partnership of members from various government ministries and offices that have an active role in the production of official statistics. The "KSH coordinates the activities of HSSz and provides guidelines and recommendations" and the "members of HSSz collect and provide data based on the National Data Collection Program (OSAP)." Some HSSz members have registers, but not all of these serve statistical goals or are a part of official statistics. The KSH maintains an address register and a register of economic organizations and is currently developing a population register.

- **Italy** has a population register that allows the production of statistical indicators on the demographic characteristics of individuals and households (e.g., age, gender, civil status, household composition), employment status, and educational attainment. Business statistics are produced from a Business Register (BR) which integrates different administrative files and the Extended BR which integrates administrative files and business surveys. In addition, the NSO is working on thematic registries on work, education and training, and persons with disabilities.

- **Latvia** has around 200 state- and municipality-held databases and registers, which are accessible to the Central Statistical Bureau (CSB) and used for the production of official statistics in accordance with the Statistics Law (Article 15). Latvia has an Official Statistics Program that specifies the official statistics to be collected, the department responsible for the production of certain statistics, and the data sources to be used. The CSB produces a large part of official statistics, but other departments (e.g., Centre for Disease Prevention and Control) are responsible for official statistics in their domains of responsibility (e.g., causes of death).

- In **Mexico** the National Institute of Statistics and Geography (INEGI, Mexico's NSO) is responsible for coordinating the NSS and regulating information collection, processing and dissemination. The NSS is comprised of "State Units" (or government ministries) that have the authority to conduct statistical activities or maintain registers on relevant information. The NSS is divided into four information subsystems that span the domains of (a) Demographic and Social, (b) Economic, (c) Geographical and Environmental, and (d) Government, Public Security, and Justice. The NSO also prepares the technical and methodological standards for statistical activities in the information subsystems of the NSS.

- The **Netherlands** has a national population register that is part of a system of base registers. "The population register is maintained by 350 Dutch municipalities and access

is coordinated by a dedicated government body." Statistics Netherlands has legislated access to all registers for statistical purposes, and routinely uses several hundreds of register sources from dozens of register holders, which are usually government bodies.

3. **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.

- **Canada**. Statistics Canada's role in linkages of administrative data: (a) supports the design, maintenance, evaluation, research and redesign of ongoing data collection and methodological studies, (b) provides statistical information in aggregate or anonymous format in support of research studies, and (c) addresses and mitigates the inherent privacy intrusive nature of these activities. Both Judiciary and Parliamentary bodies have recognized the legal authority, in accordance with the *Statistics Act*, of Statistics Canada to perform data linkages in a manner that minimizes privacy intrusions, while under the governance and authority of Statistics Canada's Chief Statistician, and requiring informing the public of such usage.

- **Estonia**. The NSO of Estonia has coordinating roles to take on large-scale data linkage across domains and sources.

- **Hungary**. The KSH has a coordinating role and the legal entitlement to use personal identification numbers for statistical purposes, but faces "both legal and technical barriers" in linking administrative data across domains and sources. "In practice, KSH relies heavily on other organisations, and has little effect on how public agencies collect data and build registers. At the end of the day, KSH remains a simple user of admin data with not much influence on quality."

- **Latvia**. "In general, the linkage of different data is possible and allowed for official statistics purposes, according to the Statistics Law, and is part of the usual practice of the CSB." The CSB of Latvia works with data providers to ensure the data quality of the registers, which is generally sufficient for the production of official statistics. In addition, the CSB has the "knowledge to link data to yield innovative statistical products tailored to the specific requirements of decision-making processes." The CSB has a leadership role in providing other institutions in the NSS with anonymized linked data for statistical purposes, based on a legal basis for the data that are needed.

- **Mexico**. The NSO of Mexico is granted autonomy by the Constitution, which is "the foundation of public trust in the quality of the statistical and geographic information" that it provides. The NSO has developed an information infrastructure that facilitates data sharing across the NSS. Legal provisions establish that the "data used to produce statistical information are strictly confidential and may not be used for any other purpose. The information shared or published must always be aggregated or anonymized to protect the privacy of individuals."

- **Netherlands**. The legal access to government administrative registers provides Statistics Netherlands (CBS) a "unique position" to leverage data from across the NSS for information purposes. However, the administrative data that CBS receives is not always error free, and "lot of effort is still put into automated and manual editing to improve quality (e.g., completeness and consistency). Data sets being used, methods and approaches are published on the CBS website to promote transparency and enhance trust," and the "CBS fully observes all applicable privacy and security requirements." The CBS is "involved in many government-wide initiatives to improve the Dutch (government) data landscape," such as the creation of a Federated Data System. The CBS also provides access to data or additional services in preparing and linking data to agencies, municipalities, and ministries for planning and policy studies.

4. **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 the coordination of data linkage activities. An enhanced role would support: (a) interoperability, (b) data curation, and (c) efficient use of resources and exchange of

information in the NSS. The prevailing theme in the responses was that NSOs are well-positioned to be **user gateways** for access to linked data.

- **Canada**. An area for future work is how to bring in new technologies and platforms (e.g., open-source data) that would allow NSOs to take the lead in working with virtual environments when current linkage platforms might no longer be optimal.
- **Estonia**. There are "further possibilities in sharing and linking data for other than statistical purposes, for instance to become a user gateway to the different sources of data across the public (and possibly private) sector."
- **Hungary**. "Building a population register is an ongoing project at KSH. There is an intention to produce experimental demographic statistics in 2025 based on the first version of a population register. Once finalized, it will open many doors for diversifying population statistics via data linkage across domains."
- **Italy**. "Our NSO can and must coordinate the main data providers (public and private) in the context of the interoperability of the IT systems of individual administrations."
- **Latvia**. The NSO can position itself as the "national analytical and competence center responsible for administrative data linkage and providing access to the linked data" within a secure environment. The technical expertise of the NSO position is to be the competent body to fulfil roles on data curation and confidentiality specified in the *Data Governance Act* (Article 7.4) of the European Union. "The incorporation of new functions and an expanded role of the CSB can occur upon the Cabinet of Ministers making the necessary decisions, including budgetary considerations."
- **Mexico**. "Data governance and stewardship should establish data sharing standards, roles, and responsibilities at the National Statistical System (NSS) level to minimize duplication of efforts and take advantage of synergies and the availability of administrative and new data sources, guaranteeing quality, interoperability, security and statistical confidentiality." These enhanced roles are needed for the "efficient use of resources" in the NSS and to "open the possibility of gaining insights into more complex problems that still need to be addressed by official statistics."
- **Netherlands**. "A potential new service is that of trusted third party. In one case, involving five institutes that have a role in the criminal law data chain, CBS already acts in such a role. In other similar cases, e.g., related to energy transition, similar constructions are considered. Often, the legal situation and institutional arrangements complicate matters. At the European level, the new EU Data Governance Act and the Common European Data Spaces may offer additional new opportunities for official statistics."

5. **Data linkage for information needs**. The questionnaire asked respondents to provide an example of major information needs in their country that have been solved by linking data. The question asked respondents to specify whether the information need was problem-driven or opportunity-driven. Most NSOs have used data linkage for problem-driven needs related to operational challenges (e.g., response burden) but it is clear from the responses that data linkage can also be opportunity-driven in that it provides more timely statistics and insights on phenomena that cannot be observed with a single source of data.

- The NSO of **Canada** has collaborated with agencies in the province of British Columbia to produce an analytical file on persons who experienced fatal and non-fatal opioid overdoses. The data file was created with a data linkage environment developed at Statistics Canada (the SDLE, described below) and integrates administrative data from (a) the provincial coroner, (b) contact with various health care services, (c) encounters with the justice system, (d) employment, income and social assistance, and (e) immigration. The data file provides disaggregated data on the socioeconomic conditions of persons who experienced an opioid-related overdose and has been used to develop insights into the heterogeneity within this population and the risk factors of overdose.
- The NSO of **Estonia** reported that "data linkage is the most valuable service we have – it is faster and more flexible than official statistics." The NSO has a dedicated team for experimental statistics who are tasked with looking for new ways to describe economic, social, and environmental phenomena using data from different sources. The NSO has conducted problem-driven data linkages to monitor the mobility of persons during the

COVID-19 pandemic and analyze the labour market situation of Ukrainian refugees. "But more and more, data linkage is also opportunity-driven" and responsive to data needs in the public and private sectors. "For instance, we have been approached by business associations to get data describing the metrics for a certain business sector (micro- and small companies, defence and space industry, companies providing cybersecurity services, etc.)."

- Data linkage for information needs at the NSO of **Hungary** has been problem-driven as it "provides multiple solutions to the problem of publishing instant and precise demographic statistics." For example, data linkages have been used to adjust for coverage errors, decrease the cost of the census, reduce response burden, and increase the data quality of surveys. Hungary also notes that "advances in IT and new methods provide new opportunities to process data from multiple sources."

- The NSO of **Italy** has developed the Italian Survey on Income and Living Conditions (EU-SILC). The income data collected through interviews were integrated with data from many administrative sources for the final determination of the disposable income of individuals and families. The use of integrated data from administrative sources and a microsimulation model has been used to determine the taxes and social contributions paid by individuals, which, added to disposable incomes, constitute gross incomes. This methodological approach was adopted to reduce response burden and increase the precision and robustness of estimates.

- In 2021, **Latvia** debuted a register-based Population and Housing Census – data from a register of population estimates were linked with data from other registers and databases. This linkage was problem-driven as it used pre-existing administrative data to replace the costly collection of census data by household enumeration and it was opportunity-driven in that pre-existing administrative data can be used to provide more frequent (annual) estimates of census variables.

- The NSO of **Mexico** has developed the Quarterly Exports by Federal Entity (ETEF) which is a source of short-term statistics on economic activities in the 32 states of the country. The ETEF is linking data from economic censuses, manufacturing surveys, administrative records, foreign trade and agricultural statistics. The Statistical Business Registry of Mexico (RENEM) is used as the connecting axis. The ETEF provides information on the state-level contributions to exports.

- The NSO of the **Netherlands** conducted its last census of population based on direct enumeration in 1971 and has since moved to a register-based census. "Probably the most important example of data linking is the System of Social Databases (SSB) that uses the Dutch population register as its backbone and links all kinds of data sources relating to persons and households, including both administrative registers and survey-based data sets, to this backbone." The SSB is opportunity-driven in that census tables are based on data already available in the system and "census statistics are derived from the SSB with not much more than the push of a button." The SSB has recently been expanded to examine social networks, which adds information on relations (e.g., family, school, work) between people.

6. **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. As detailed below, 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.

- Statistics **Canada** has developed the Social Data Linkage Environment (SDLE), one of several linkage environments. For various projects, the SDLE has been used to link person-level records from 160 sources from across the domains of health, justice, education and income. Microdata linkage at Statistics Canada must adhere to the Directive on Microdata Linkage, which is designed to ensure that the public value of each linkage outweighs any intrusion on privacy that it represents.

- Statistics **Estonia** has the legal right for data linkage across sources and uses an open-source software and ecosystem solution (X-Road) that provides unified and secure data

exchange between providers of data. The NSO has established the IT infrastructure to process and link data.

- In **Hungary**, administrative data can be used for official statistics without restriction, under law, but this still requires cooperation agreements with the organizations that control the data. The NSO has an IT infrastructure for bringing the external data into its database (KARÁT), transforming the data (ADAMES), and storing the original and cleaned data.
- In **Italy**, "a national data strategy has been outlined with the objective of making the databases managed by individual Public Entities interoperable. At the core of this strategy there is the development of the National Digital Data Platform (PDND), which is a technological infrastructure available to all Public Administrations (PA). This platform enables the interoperability of information systems and databases of entities and managers of public services, embodying the **once-only principle**, which aims to reduce the statistical burden on citizens and businesses. The platform is designed to support citizens, businesses and public decision-makers by simplifying administrative procedures and enhancing the Italian 'information capital' of public administrations."
- The Statistics Law mandates the NSO of **Latvia** to participate in the development of administrative data, promote the use of administrative data for official statistics, to coordinate the standardization of data, and take responsibility for data stewardship of the state information system.
- The NSS of **Mexico** uses a shared information infrastructure (catalogues, statistical and geographical registries and methodologies) that facilitate the integration or linkage of information from different production processes. The NSO is currently experimenting with data science projects that use data lake architectures to automate data workflows and have the potential to increase the efficiency of data-linking tasks.
- In the **Netherlands**, linking data is facilitated through unique identifiers that are available in most registers. "As mentioned above, in the domain of social statistics the SSB is in place, which includes a lot of protocols, methods and tools to ensure robust data access and high-quality data. For business statistics, a new system is being developed that supports routinely linking URLs to companies in the statistical business register. Other business surveys and administrative sources are already routinely made available through their statistical business registers IDs."

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

- Value of unique identifiers:
    - "The lack of a unique identifier makes linking difficult. It is necessary to test more probability linking methods, however, these methods can also result in incorrect linking due to the different reference populations" (**Hungary**).
    - A common ID code 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 than is available with surveys or the census (**Latvia**).
    - Using a single registry as a connecting axis was a valuable piece of information infrastructure that made it possible to generate short-term statistics on quarterly exports and monthly indicators (now-casting) of manufacturing activity (**Mexico**).
    - "Linking data sets referring to persons is also facilitated through the Citizen Service Number (BSN) that is used as unique identifier in most registers." This does not mean that everything goes automatically. Regular contacts with register holders are necessary to make sure that all relevant microdata needed for official statistics arrive on time and with accompanying metadata (**Netherlands**).

- Need for data sharing agreements:
    - To mitigate privacy concerns when linking personal level data from different sources and facilitate computation on sensitive data, Statistics Canada has partnered with other federal government organizations to explore the use of Privacy Preserving Record Linkage (PPRL). Protocols used in the PPRL allow for additional privacy protections of microdata. The analytical work can be computed on the linked dataset with necessary masking and there is no need of transferring potential sensitive data from one organization to another. This field can have future implications for data linkage and integration (**Canada**).
    - "For now, we rely heavily on data obtained from registers and official statistics, which does not always cater to the specific needs of the users. For better and more diverse data, there is a need also for the data held by the private sector." The access to this data is restricted because of legal and trust issues. "The use of privacy-enhancing technologies that could facilitate the social acceptance and trust from the private sector are not yet very cost-efficient" (**Estonia**).
    - "Obtaining administrative sources is still too slow and difficult, the tools to formalise relationships with other administrations, such as agreements, protocols, conventions etc., often take too long to satisfy the needs for data transfer, analysis and valorization. The current legislative framework to access and use administrative data apparently is not sufficient." (**Italy**).
    - "Access to sources like administrative records depends on establishing agreements with third parties, and there is a risk associated with losing the data flow" (**Mexico**).
    - "The practical implementation of data access has led to regular contacts and contracts between CBS and register holders. In these contracts, duties and rights of both parties are mentioned. It is, e.g., important to know when new data will arrive and in what format. This way the new data can quickly be included in the production systems of CBS. In the contracts, contact persons are mentioned so that the contacts are streamlined and questions to register holders are only asked once" (**Netherlands**).

8. **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 on-going international activities on the technical aspects and challenges of data linkage that are important initiatives.

- "At the European level the use of unique identifiers in agricultural statistics is a recurring theme , and an international project on this subject has just been launched" (**Hungary**).
- "Projects whose objective is the use of non-probabilistic sources for statistical purposes. Data linking is one of the key tools, e.g., when non-probabilistic data take advantage of the representativity with respect to a population given by a sample or statistical frame. This is an issue, for instance, in case of mobile phone data" (**Italy**).
- The HLG- Applying Data Science and Modern Methods Group and Supporting Standards Group were cited as important ongoing collaborations as was the UN Committee of Experts on Big Data and Data Science (**Mexico** and the **Netherlands**).