United Nations ECE/CES/2024/5/Add.1

Economic and Social Council

Distr.: General
14 June 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

Outcome of the electronic consultation on the in-depth review of linking data across domains and sources

Prepared by the Secretariat

Summary

This document presents the main results of the electronic consultation on the in-depth review of linking data across domains and sources, that was conducted among the members of the Conference of European Statisticians (CES) in April–May 2024.

The in-depth review of linking data across domains and sources was carried out by the CES Bureau in February 2024, based on a paper (ECE/CES/2024/5) prepared by Canada.

The Conference will be invited to endorse the outcome of the in-depth review of linking data across domains and sources.

I. Introduction

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

2. The Bureau carried out an in-depth review of linking data across domains and sources in February 2024 based on a paper by Canada (document ECE/CES/2024/5).

3. The UNECE Secretariat conducted an electronic consultation in April–May 2024 to inform all CES members about the in-depth review of linking data across domains and sources and provide an opportunity to comment on its outcomes.

4. The following 26 countries and organizations replied to the electronic consultation: Austria, Belgium, Canada, Chile, Costa Rica, Ecuador, Finland, France, Hungary, Ireland, Japan, Kazakhstan, Latvia, Malta, Mexico, Netherlands (Kingdom of the), New Zealand, Poland, Portugal, Slovenia, Sweden, Switzerland, Türkiye, the United Kingdom of Great Britain and Northern Ireland, the United States of America and the Organisation for Economic Co-operation and Development (OECD).

II. Outcome of the Conference of European Statisticians Bureau discussion in February 2024

5. The Bureau carried out an in-depth review of linking data across domains and sources at its February 2024 meeting.

6. The Bureau considered that 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 Bureau also noted that the proposal to develop a road map would be a constructive way forward in international work on this topic. Detailed comments made by the Bureau at the February 2024 meeting are available in document ECE/CES/2024/5 (para. 60)

7. The following conclusions were reached:

(a) The High-Level Group for the Modernisation of Official Statistics (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.

III. General comments received in the electronic consultation

8. Among the 26 countries and organizations that replied to the electronic consultation, 17 provided general or detailed substantive comments, and 8 did not have any comments.

9. Many countries praised the quality of the in-depth review and underlined the importance of linking data across domains and sources. **Belgium** noted that the paper gives a good overview of the issues related to linking data and the challenges experienced by the

national statistical offices (NSOs). **Chile** commented that linking data is of crucial importance for the current context of statistics, so its further study is particularly relevant. In **Finland**, linking data across domains and sources is very important and helps providing information about groups of persons who are not included in the population information system (i.e. who do not have a legal domicile and who are therefore not covered by the official population statistics), like asylum seekers, temporarily workers in the country or cross-border workers. In **France**, the statistical system has been matching individual data for many years, enriching survey data with administrative data, and matching administrative data with each other. Continuing along this path by making progress in data collection and matching techniques is one of the priorities of the modernization of statistical data production. **Ireland** noted that the review provides valuable insights as spines and registers of data are built in the Central Statistics Office (CSO). **Latvia** appreciated the work done and supported the outcomes of the in-depth review, noting that data linking is a way to develop innovative statistical products tailored to the specific requirements of decision-making processes. **Malta** commented that the mainstreaming of data linking issues, collaboration, and innovation in data linkage initiatives are rightly emphasized throughout the document as essential methods for improving data transparency and communication with stakeholders. **Sweden** appreciated the initiative for the in-depth review on data linking across domains and sources and noted that this issue is closely related to the issue of the NSOs' role as data steward, interoperability issues and access to privately held data. **Switzerland** noted that linking data is an important task that adds real value to official statistics and added that the difficulties encountered with the problems of non-response and/or coverage can be overcome (at least partially) by linking the data. The **United Kingdom** welcomed the outcome of the review, noting that collaboration with experts across the field has helped in the development of linkage work within the organization and international collaboration remains an important area for learning, improving and developing methods and application. The United Kingdom added that the horizontal issues mentioned in the paper are a recognized common challenge and discussions and progress in these areas would greatly aid data linkage development as a tool for future research and statistical production.

10. **OECD** welcomed the CES work on data linking across domains and sources and endorsed the report, noting that data linking is key to better understand the complexity and interconnections of economic, social and environmental developments, and that this will help analysts undertake multidimensional analyses but also fill in some important data gaps (e.g., on gender).

11. Several countries also made the following comments on various issues discussed in the in-depth review paper.

12. Concerning **privacy and data protection**, **Chile** noted that it is highly relevant to delve into recommendations related to ensuring data privacy. NSOs generally take precautions regarding the anonymization of records from individual sources, but when linking data, new opportunities for re-identifying a reporting unit emerge due to the incorporation of new variables from other sources. For **Mexico**, as NSOs take on their new role as "user gateways", it is important to support the accessibility and democratization of information to enhance its usability and impact, at the same time data privacy must be maintained and advocated to ensure that sensitive information remains protected. **Switzerland** recommended that the necessary measures must be taken to guarantee data protection so that the data linking approach is accepted by the public, to avoid possible risks.

13. **France** noted that the document rightly emphasizes issues of **social acceptability**, considering that in linking data each NSO is making progress according to its possibilities, needs, legal and institutional context and the level of acceptance of such processing by the population.

14. With regard to **economic costs** associated with data linking, **Mexico** commented that for large volumes of administrative records the costs should be considered in advance, in addition to the technological capabilities needed to manage and process the data effectively. The cost can be a significant barrier for countries and may exacerbate existing capacity gaps. On the other hand, **Switzerland** noted that the data linking approach can lead to a reduction in the burden on respondents and in data collection costs, by increasing efficiency and helping

to overcome (at least partially) the difficulties encountered with the problems of non-response and/or coverage.

15. **Mexico** also commented that, although a systematic approach is necessary for data linkage projects, such an approach must remain **flexible** enough to accommodate new alternative data sources that may differ from those presented in the proposal. Given the rapid pace of technological advancement, the original proposal could become obsolete if it does not adapt to future developments.

16. **Netherlands (Kingdom of the)** emphasized the importance of the **businesses register, the jobs register**, which allows Statistics Netherlands to link people and companies (the so-called LEED, linked employer employee database), and the system of basic registers that enables linking many different registers.

17. **Switzerland** noted that, if a unique identifier is lacking, care must be taken when using probabilistic data links because they can lead to a deterioration in the **quality** of the results.

18. **OECD** also commented on **quality** – a key challenge in data linking is to reconcile data of different qualities and coverage and which may provide an inconsistent view of a single phenomenon. OECD fully supported the recommendation to document and communicate clearly on how this challenge has been addressed to users. This will allow them to get a clear understanding of what is behind the statistics so that they can make an informed decision on how to use them. OECD also noted that a common ID code or single registry are at the moment the first best to undertake data linking (see para. 47 of the in-depth review paper). However this is not always possible and can entail financial costs. It would be useful to investigate further how alternative data could help, keeping in mind the technical difficulties of exploiting those data and the fundamental quality principles associated with official statistics.

19. The **United States** noted the key recommendation of the review that NSOs should have a coordinating role in data linkage, operating as "user-gateways" for linked data. In the United States, the forthcoming National Secure Data Service (NSDS) may be able to serve in this capacity in the future, while at an agency scale, the U.S. Census Bureau is implementing a similar solution by building a Business Ecosystem (BE) to serve as a home for linked data from multiple sources residing in geospatial, person, business, and job frames.

20. The **United States** also suggested adding to the second area of discussion ("the complex information needs of policymakers are a driver of linking data across domains and sources") that the rise of **artificial intelligence (AI)** is also potentially driving data linkages (e.g., in pre-training foundation models or in later stages such as fine-tuning or retrieval-augmented generation applications). AI will understand the world better if we can provide linked data.

Annex

Detailed substantive comments received from electronic consultation

As a result of the consultation, substantive comments, requiring possibly an update of the in-depth review paper, were submitted by Austria, France and Hungary. Those comments are presented in the table below.

Country / Organization	*Comments*
Austria	Para 23 – A “centralized system” is described as a disadvantage in terms of user-relevant statistics, because NSOs detached from the policy discussion allegedly lack subject matter expertise. Maybe a concrete example would be helpful for understanding this statement. Furthermore, a close relationship between production of statistics and politics may be a problem for producing independent statistics. Para 34 – Compared to Poland and Canada, a more concrete example of problem or opportunity driven challenges have been solved by linking data would be helpful (reduction of poverty in (c) is only described very generally). Para 37 – As described in the document, using linked data only for statistical purposes should be regulated by law, in order to ensure confidentiality. It would be important for the countries that only took part in the review (Poland, Russian Federation) to provide a description of the legislative basis of data linkage within the NSS too (as is the case for countries that took part in the survey). Para 41. (a) – Is statistical matching meant by the term “probabilistic linkages”? Although we agree – as it is described in the document – that further work in testing these methods, is helpful we like to propose to refer to research which has been carried out already in this field. (cf. Usage within the ESS https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-RA-13-020, “Statistical Matching: Theory and Practice” in d’Orazio et al. (2006) https://www.wiley.com/en-us/Statistical+Matching%3A+Theory+and+Practice-p-9780470023532) Para 60 (l) – Statistics Austria provides data linkage predominantly for register data. Certified users have access to linked data via the Austrian Micro Data Centre (AMDC). Maybe this should be mentioned in the document, similarly to the case of the UK as in Section VIII. 60 (l).
France	Para 60 (e) – We have serious reservations about the wording at the end of paragraph 60(e), which could call into question the protection of data collected for statistical purposes. The "rebalancing" must not take place through the transfer of individual statistical data to administrations, as this is prohibited by our national law. In view of this ambiguity, it seems to us that the paragraph could be withdrawn, or at least the last sentence: we would confine ourselves to pointing out a potential difficulty to which there is no single answer.
Hungary	In the document, legal implications of data privacy issues are left somewhat vague.

Country / Organization	*Comments*
	Paras 9, 10, 48, 53, 59 (a) – The document talks in **para 9** about NSSs and NSOs having "the opportunity to bring […] data together". While these organizations might have broader rights to link data and use unified datasets, there are still legal constraints over the types of linkages and usage of linked data that apply to NSSs/NSOs and thus need to be considered. Para 9 goes on stating that there needs to be a "once-only principle" for collecting data once and reusing data later by NSSs/NSOs. This could contradict the principle of purpose limitation of the GDPR if stringent guidelines are not set up for reusing data and informing the data providers about such data usage. This need for guidelines should be highlighted in this section. The same should be emphasized in **para 10** as well, where examples of reused administrative and commercial data are listed, without any hint on the steps taken to inform data providers about further usage of their data collected originally for administrative and commercial reasons, respectively. Similar considerations should be added to Point 40 about the requirements of facilitating data linkage; and to **para 48** where public acceptance is listed as a dimension of data ethics when re-using administrative data, but transparency is emphasized in less detail. In **para 53**, creating data standards for re-using data is proposed, but again, the need for informing data providers about the future utilization of their data is not emphasized. The listing in **para 59 (a)** of section "Conclusions and recommendations" should include confidentiality and transparency as well, together with "ensure[ing] sound implementation and policy relevance". Para 5 – The listing of "new technologies and tools (such as artificial intelligence (AI) and machine learning (ML)" might need some further consideration whether ML and AI are actually considered to be distinct entities. Para 41 (a) – The reference to probabilistic linkages might need to include information about Machine Learning based solutions that could further enhance linkages in the presence of not perfectly matching identifiers (and in the absence of unique identifiers). There might need to be a short listing of the potential advantages and risks added, both about probabilistic linkages and (if applied) Machine Learning based solutions. Para. 56 – The text might want to emphasize the potential development of reusable, multi-purpose indicators, together with their data privacy challenges. **"Annex" (ECE/CES/2024/5/Add.1)** – We would appreciate if you could make changes in the paragraphs concerning Hungary in the Annex based on the suggestions provided in a separate document sent to UNECE.