UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

ECE/CES/BUR/2017/FEB/2
7 February 2017

CONFERENCE OF EUROPEAN STATISTICIANS

For discussion and recommendations

Meeting of the 2016/2017 Bureau
Geneva (Switzerland), 14-15 February 2017

Item II (a) of the Provisional Agenda

In-depth review of data integration

Prepared by the secretariat and the participants and project leader of the HLG-MOS 2016 Data Integration Project

The present note is an in-depth review of data integration, based on the experience from the HLG-MOS 2016 Data Integration Project.
The conclusions and recommendations from the review are presented in section VI. The note was presented to the CES Bureau for discussion and recommendations.

I. Summary

1. This review describes the experiences gained from the experiments in data integration undertaken in the 2016 Data Integration Project in which the national statistical offices of Brazil, Canada, Colombia, Hungary, Italy, the Netherlands, New Zealand, Poland, Serbia and Slovenia participated. Several types of data integration were considered: survey and administrative sources, survey and new data sources (including Big Data), traditional sources with geospatial information, and integrating data for validating official statistics. Under these types, several experiments were conducted resulting in the identification of opportunities, challenges and issues, and leading to many recommendations and lessons learned.

2. Amongst the many opportunities offered by data integration are new or timelier statistics that are cheaper to produce, reduce the response burden and are potentially of improved quality. The main challenges identified were: stability of data sources, skills and technology needed, conceptual differences, metadata needs, quality management, effective partnerships, moving from testing to production, setup and maintenance costs, avoiding duplication, and public perception regarding privacy. With ever more data sources becoming available and increased capacities of IT and data infrastructure, the need for integrating different sources is growing.

3. Many different types of data and different sources can be integrated. Therefore there are no common recommended methods for integrating all types of data. However, there are several standard processes across different types of data integration. The data integration projects could follow a similar series of steps, such as identifying needs and clarifying business requirements, identifying partners, selecting potential data sources, considering methodologies and quality, analysing costs, benefits, risks, and obtaining the data and the required tools, skills and resources. When the experiments are done, the results should be assessed, and the methods and approaches refined to develop them into a repeatable production solution.

4. Developing quality frameworks for integrated data and addressing metadata requirements are important to maintain the value proposition of official statistics produced from integrating data sources. By its nature, various types of partnerships need to be built for data integration, leading to additional challenges for NSOs. Development of strong governance frameworks for managing data integration projects is important. In addition to essential substantive statistical skills, new IT skills, computer applications and hardware and the skills needed for obtaining, negotiating and communicating the use of external data sources should be acquired.

5. The various data integration experiments lead to a large number of recommendations, for example:

- get high level management support and let management of statistical offices know about the project and its goals and results
- keep your objectives clear and clarify importance of the project, specify the characteristics of the problem to be resolved
- good collaboration with partners/data providers: be clear about the data requirements, take into account the common goals of institutions (collecting data only once, reducing unnecessary expenses), set up agreements
- collaborate with users of the data
- consult with other experts, find out more on possible methods and solutions in a practical sense; exchange personnel to gain experience and learn from good practices
- consider international recommendations and standards for statistics
- share results
- start the work on a sample, look at the data before commencing, pertain patience and persistence, have in mind a wide range of solutions

- measure data quality

6. The HLG-MOS data integration project continues in 2017. **The Bureau members are invited to:**

- **comment on the main conclusions of the in-depth review** (Section VI)
- **nominate their experts to join the 2017 HLG-MOS Data Integration Project**, particularly in the production of the synthesis report. The tasks include drafting and finalising guidelines and recommendations, and making them available as an on-line *Guide to Data Integration* for the statistical community
- **advise on how to further develop the work on data integration** after the current project ends in December 2017.

II. Introduction

7. The Bureau of the Conference of European Statisticians (CES) regularly reviews selected statistical areas in depth. The aim of the reviews is to improve coordination of statistical activities in the UNECE region, identify gaps or duplication of work, and address emerging issues. The review focuses on strategic issues and highlights concerns of statistical offices of both a conceptual and a coordinating nature.

8. The CES Bureau selected data integration in October 2015 for an in-depth review at its February 2017 meeting. The current paper provides the basis for the review by summarising the international statistical activities in the selected area, identifying issues and problems, and making recommendations on possible follow-up actions.

9. Data integration provides the potential to produce more timely, more disaggregated statistics at higher frequencies than traditional approaches. In 2015, the High Level Group for the Modernisation of Official Statistics (HLG-MOS) recognised that official statistical organisations were challenged by the capacities needed to incorporate new data sources in their statistical production processes. This resulted in the 2016 HLG-MOS Data Integration Project. The review is based on experiences and lessons learned by the participating countries in the project.

10. The paper outlines the types of data integration covered in Section III, a general framework for data integration and the broad approach for the data integration projects in Section IV. Section V summarizes the issues, opportunities, challenges, risk mitigation, recommendations, skills and resources needed. The last section provides conclusions and summary recommendations, and a proposal to the CES Bureau. A summary of each experiment proposal is available from the 2016 HLG-MOS Data Integration Project wiki at: http://www1.unece.org/stat/platform/x/HQazBw.

III. Scope/definition of the statistical area covered

11. The aim of the 2016 HLG-MOS Data Integration project was to gain experience that would allow to develop general recommendations and guidance for data integration and a related quality framework. The project pooled resources in joint practical activities.

12. There are many possible types of data integration using many combinations of data sources and modelling approaches. It would not have been feasible to cover all types of data integration in a single year project. Therefore it was decided to focus on experiments covering four main types of data integration, as described below (in brackets is given the country that proposed and participated in the project, more countries may have participated in the experiment when it was carried out).

13. The selected types and experiments undertaken were:

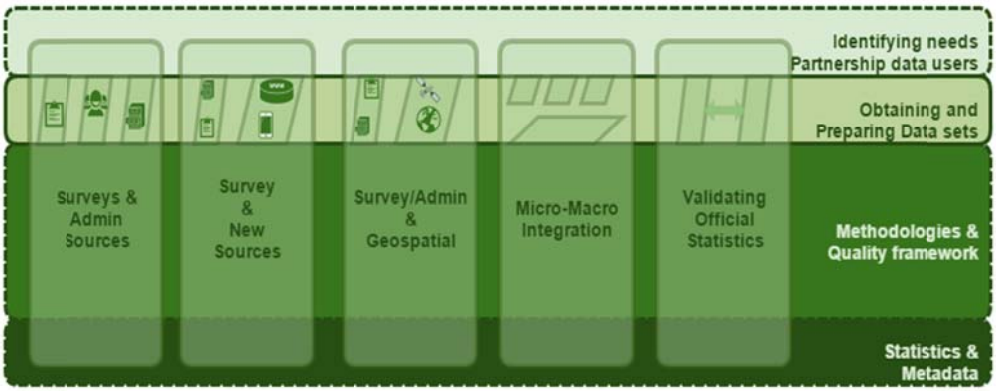
1. **Integrating survey and administrative sources**
    a. Integrating information on education based on geographic location of schools (Colombia)
    b. Integrating potential information sources for producing statistics on job vacancies (Hungary)
    c. Linking the Statistical Register of Employment and the Labour Force Survey (Slovenia)
2. **Integrating new data sources, such as big data, and traditional sources**
    d. Web-scraping strategy for official statistics – case study of New Zealand and survey of other countries' experiences (New Zealand)
    e. Integrating web scraped data for the compilation of price statistics (Hungary)
    f. Integrating scanner data and web-scraped price data to enable internationally comparable price measurement (New Zealand)
    g. Estimation and comparison of price indices from different big-data sources across countries (New Zealand)
    h. Integrating potential information sources for producing statistics on job vacancies (Hungary)
3. **Integrating geospatial and statistical information**
    i. Integration of spatial objects used in statistics and geodesy - „The 10 Level Model" for harmonizing the reference frameworks of statistics and geodesy (Poland)
    j. Analysis of existing models of integrating geospatial and statistical information (Colombia)
4. **Validating official statistics using data from other sources**
    k. A comparative analysis of income data from New Zealand Income Survey and administrative data (New Zealand)
    l. Linking the Statistical Register of Employment and the Labour Force Survey (Slovenia)

14. The 2016 HLG-MOS project had identified a fifth type of data integration, "**Micro-Macro integration**" but no experiments were undertaken in this area.

15. Some activities related to more than one type of data integration. There were also cross-cutting activities that aimed to obtain and prepare data sets and to

synthesise experiences with respect to partnerships, methodologies, quality frameworks and metadata requirements. There was also an activity across the project to identify and test the tools used for Data Integration in the various experiments listed above. Figure 1 gives an overview of the structure of the project.

Figure 1: 2016 HLG-MOS Project Structure



16. **This review is also based on the information gained through the 2016 HLG-MOS project and experience from other data integration work already completed or in progress.** It explores issues and lessons learned with the aim to develop guidance to support and advance data integration activities in official statistics. The review highlights some of the issues that may need to be considered in order to undertake successful data integration projects.

17. The following statistical offices have contributed to the project: Brazilian Institute of Geography and Statistics (IBGE), Central Statistical Office (CSO) of Poland, Hungarian Central Statistical Office (HCSO), Italian National Institute of Statistics (Istat), National Administrative Department of Statistics (DANE) Colombia, Statistical Office of the Republic of Serbia (SORS), Statistical Office (SURS) Slovenia, Statistics Canada, Statistics Netherlands (CBE), Statistics New Zealand, Eurostat and the United Nations Economic Commission for Europe (UNECE).

IV. Overview of the experiments on data integration

18. This section provides a brief overview of a framework for the data integration projects, and a description of projects that were implemented on different types of integration. A summary of each experiment proposal is available from the 2016 HLG-MOS Data Integration Project wiki at: http://www1.unece.org/stat/platform/x/HQazBw.

19. Practical work with experiments starts with data. If there are no suitable data sets that can be shared between organisations, a sample or synthetic datasets can be obtained or developed to share methods, tools and experiences for integrating data. The Ireland's High-Performance Computing Centre's Sandbox was available for processing and testing of large (big) data sources. It is a relatively open environment that is not designed for dealing with privacy-sensitive or otherwise confidential data. Various statistical, methodological, legal and ethical issues need to be considered for data integration but use of data sets only for experimentation makes this easier. As suitable methods, processes and tools are developed in a collaborative way, they can be moved to the secure environments of individual organisations for further testing on real data.

A. Framework for data integration

20. There are many possible types of data integration and for each, many possible combinations of data sources and modelling approaches. Data integration can be done at the micro level, at the level of a common denominator, at the aggregate (macro) levels, through modelling approaches or a mixture of these.

21. Five common types of integration are:

(a) administrative sources with survey and other traditional data;

(b) new data sources (such as big data) with traditional data sources;

(c) geospatial data with statistical information;

(d) micro level data with data at the macro level; and

(e) validating data from official sources using data from other sources.

22. The new opportunities arising from multiple data sources and the problems with using surveys as the most common approach to generating official statistics led the director of the US Committee on National Statistics, Connie Citro, to state that "We must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources"(Citro, 2014)[1]. The challenge is therefore to integrate diverse sets of inconsistent data and to produce stable outputs with often unstable, ever-changing inputs. Instead of trying to produce the best possible statistics from a single survey, official statistics need to try to find the best combination of sources to deliver the indicator/statistics that best satisfy the users' needs.

23. Survey data, administrative data, big data and other non-traditional sources should be considered. Although there have been a number of attempts to integrate various data sources to produce statistics, **no generalized methodology or quality framework exists**. To address this urgent and complex common challenge, the HLG-MOS project aimed to pool resources to make a start with a more systematic approach towards developing a new common framework for statistical production.

24. Many issues should be considered in undertaking data integration projects, such as identifying needs, building partnerships, obtaining data; finding appropriate modelling approaches, and managing quality, risks, comparability and metadata requirements. A set of broad topics was developed under the project to organise the issues to consider. These are:

- Business requirements
- Opportunities
- Challenges
- Risk mitigation
- Standard processes
- Recommended methods
- ICT considerations
- Quality
- Standards
- Metadata requirements
- Related work in other projects/organisations
- Skills
- Resources
- Partnerships
- Governance

[1] Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*(2), 137-161.

- Promotion and advocacy
- Recommendations

25. Participants in the experiments considered these topics, generated lessons learnt and, where relevant, developed recommendations. The project aims to bring the experiences together to synthesise more general recommendations which relate to more than one type of data integration.

B. Integrating survey and administrative sources

26. Some countries already have extensive experience integrating survey and administrative data sources. There have been collaborative projects in this area, for example at Eurostat.

27. The administrative data may have existed for some time but not been used. It may be integrated using record linking or statistical matching or by using modelling approaches. It may involve pooling or combining information from multiple surveys, including surveys not conducted by the NSOs themselves.

28. There are common challenges faced in this type of integration. The quality of administrative dataset may be good enough for administrative purposes but not sufficient for statistical purposes. Transforming administrative datasets into statistical datasets may require improving the quality and dealing with conceptual differences, especially when the administrative data is planned to be used directly. In the case of surveys carried out with the use of data from administrative sources it is crucial to gather all data (from survey and administrative sources) in one database.

29. Examples of sources that can potentially be integrated are: Labour Force Surveys and social insurance register and/or educational registers, data from ministries of culture and cultural associations to produce statistics on museum attendance. There are several examples of administrative data being combined with survey data for producing indicators traditionally collected through censuses.

1. Brief description of the experiments

30. The experiments that took place were:

- Integrating potential information sources for the statistical data production on job vacancies
- Linking the Statistical Register of Employment and the Labour Force Survey
- System of consultation and geographic location of schools

31. The experiments for this type of data integration typically included the following steps:

- Definition of statistical framework and boundaries of the work.
- Select Administrative sources and a survey dataset
- Define a case using records from the administrative dataset instead of records from a survey (a smaller frame) or adding additional variables from the administrative survey (a shorter a survey questionnaire) or both.
- Design expected outputs of the integration
- Design and test of transformation of the administrative dataset into the statistical dataset. - A cleaning method has to be designed
- Design and test of the integration of administrative data and survey data.
- Evaluation of statistical outputs obtained from the integration of administrative datasets and survey datasets.

2. Related work in other projects/organisations

32. Integrating survey and administrative sources is not a new topic for official statistics. Work has been done in this area over the years by different organisations, including under the CES work programme. From the currently ongoing work the following can be mentioned:

33. UNECE Assist: Knowledge base on the use of administrative and secondary sources in statistics http://www1.unece.org/stat/platform/display/adso/ASSIST

34. ESS.VIP ADMIN Project: The project purposes are to support the EU Member States to reap the benefits (decrease costs and burden, increase data availability) of using administrative data sources for the production of official statistics, and to guarantee the quality of the output produced using administrative sources, in particular the comparability of the statistics required for European purposes. https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en

35. ESSnet project on Data Integration: The project focused on methodologies for data integration (Record Linkage, Statistical Matching, Micro integration Processing) and on statistical aspects to be considered to make those methods concretely applicable by NSIs. http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en

36. ESSnet project: Integration of Survey and Administrative Data. The project purpose was to promote knowledge and application in practice of sound methodologies for the joint use of existing data sources in the production of official statistics. http://ec.europa.eu/eurostat/cros/content/isad-finished_en

37. Eurostat (2013): The use of registers in the context of EU–SILC: challenges and opportunities. http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

38. UNECE (2007): Register-based statistics in the Nordic countries. Review of the best practices with focus on population and social statistics. http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764

39. UNECE project on Quality Indicators for the Generic Statistical Business Process Model (GSBPM). This is an on-going project aimed at developing quality indicators to monitor the quality of the statistical production process for each of the phases of the GSBPM, including the sub-phase, 'integrate data'. The project is currently reviewing and updating the quality indicators to include the use of administrative data in the production of official statistics. On-going work is available from http://www1.unece.org/stat/platform/display/QI/Quality+Indicators+Home.

C. Integrating new data sources (such as big data) and traditional sources

40. This section focuses on integrating new data sources (such as big data from mobile companies, social networks, smart sensors, satellite imagery, web pages, credit card transactions, etc.) with traditional sources of official statistics.

1. Brief description of the experiments

41. The following experiments took place:

- Web-scraping strategy case studies: Approaches to gaining web-scraped data for official statistics – case study of NZ and survey of other countries' experiences.

- Integrated big-data price measurement: Estimation and comparison of price indexes from different big-data sources across countries (New Zealand)
- Integrating web scraped data for the compilation of price statistics (Hungary)
- Integrating potential information sources for the statistical data production on job vacancies (Hungary)

42. These experiments typically took steps such as:

- Define a statistical framework and boundaries of the work.
- Design and test the integration between Internet-scraped data and traditional statistical datasets, including entity extraction and recognition and object matching. Exploring new techniques for record linkage and object matching (i.e. where an object can have a looser structure than a record).
- Design and evaluation of statistical outputs obtained from the integration/combination/fusion of Internet-scraped data and traditional statistical datasets.

43. During 2016, participants in this activity investigated current work in this area (within their own organisations and in other groups (eg ESSNET and the HLG-MOS Big Data project). The activities will continue in 2017, with the proposal to:

- Develop a practical guide to integrating survey, administrative data and big data (including case studies) based on the work being done by the participating organisations
- Encourage the involvement of other participants and projects
- Describe the steps needed using GSBPM

2. Related work in other projects/organisations

44. The related work carried out in international organizations or under other experiments of the HLG-MOS 2016 Data integration project include:

- UNECE big data, Eurostat big data
- The prices work (see section B)
- New Zealand rental index from trade me site (not quality adjusted)
- Job vacancies (Serbia)

D. Integrating geospatial and statistical information

45. This section addresses the integration of geographic information with statistical information.

46. Statistical data is almost always related to a certain physical space, like a municipality, a State, a Country a region.. Each level is useful for different actors and different kinds of decisions. Many of those decisions are conditioned by physical elements from the environment, and beyond that, they will have an impact on it. Location and amounts of natural resources, soil types, weather conditions, communication infrastructure, facilities are examples of geographic information which are indispensable elements to fully understand the figures that official statistical generates. Work in the field of geospatial classification could be a departure point to link the fields of statistics and geography.

47. Integration of geographical data into the official statistical field aims to improve the value of the statistical information that is being produced.

1. Brief description of the experiments

48. Due the complexity implied in this working package it was not expected to cover all the activities in just one year. This was just the start of the work as it is likely that future activities in this area will demand efforts that will go beyond a

single task team. The experiments undertaken for this type of data integration focused on:

- Inventory of the level of integration of spatial objects used in statistics and geodesy - „The 10 Level Model" for harmonization of the reference framework of statistics and geodesy (Poland)
- Analysing a scheme of integrating geospatial and statistical information Colombia)
- Integrating information on education based on geographic location of schools (Colombia)

49. Within the experiments, the following steps were taken:

- Review international efforts in this area (like Eurostat's GEOSTAT and GISCO and UN's GGIM) to generate synergies and improve the capacities to work in the project
- An inventory of sources and projects related to geographical information which could be relevant to the production of statistics
- Develop proposals for valuable information that could be produced by integrating statistical and geographical information
- Research on methodologies to integrate and produce relevant information products derived from the combination of both kinds of data
- Conduct experiments and pilots to assess the value of integrated statistical and geographical data to adjust the scope of the project

50. The following activities were proposed for follow-up in 2017:

- Develop a decision tree - a path of practical questions to be answered before embarking on integrating geospatial data with statistical data. This will assist organisations to assess their maturity and capability for spatial statistics.
- Examine and consider incorporating the work done by participants of the GEOSTAT 2 project and Australia to reference geospatial capabilities and data standards (including standards from the geospatial community) in the components of the components of GSBPM and GSIM.
- Examine and consider joint work with UNGGIM, Europe and Americas to obtain better results in integration of statistical and geospatial information.
- Compare and analyse the results of the surveys conducted in different regions. Conduct additional surveys based on GEOSTAT2 as required if methodologies of surveys mentioned above are different and results are not comparable. Analyse the existence of common points and divergence of approaches at the national and international level of spatial objects used in statistics and geodesy to harmonise this two reference frameworks (implementation "The 10 level model" into the lower layers of GSGF).
- Identify common risks for integrating geospatial and statistical data.

2. Related work in other projects/organisations

51. The CES Bureau reviewed "Developing geospatial information services" in-depth in February 2016. A CES seminar on this topic was held in April 2016 and the Bureau discussed follow-up to these activities in October 2016. A proposal for joint activities by statisticians and the geospatial community under the Conference will be considered under agenda item III(k) at the February 2017 meeting of the Bureau.

52. The geospatial and statistical data integration landscape is very complex, with many players globally. The Global Statistical Geospatial Framework (GSGF - UN GGIM), the Statistical Spatial Framework (SSF), merging statistics and geospatial information in EU member states (ESSnet), European Forum for Geodesy and statistics (EFGS), and initiatives such as GEOSTAT2 (Eurostat) are vital for developing a consistent and systematic approach to linking geospatial and statistical data. This is likely to take some time.

53. **Because of the complexity of this area, it would be useful to conduct a face to face or virtual sprint of key players (including Australia, Poland, Colombia, Mexico, Finland, Sweden and others to be determined. This is expected to be incorporated into a proposed workshop on geospatial and statistical standards.**

E. Validating official statistics

54. Another specific area where data integration comes into play is using data from other sources for validating official statistics. Since data integration enables identification of records from multiple data sources that belong to a single individual or unit, it can be used to validate official statistics, either through:

- use of external data sources to determine accuracy of survey results or
- use of survey results to challenge results from alternative data sources of providers of statistics.

55. There have been cases where other sources are seen as comparable to official statistics, and when they differ, the official statistics have been challenged. One example from the United Kingdom shows how the distribution of businesses listed in the "Yellow Pages" telephone directories was compared to the coverage of the statistical business register (http://www.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf). A further example concerns comparisons of inflation figures from MIT's Billion Prices Project against official price indices. These examples show that "other sources" are reaching a level of credibility that challenges the role of official statistics. Market researchers are struggling with similar issues. Their business is already under increasing pressure from cheap or free internet panels.

56. One thing is clear. These "other sources" are here to stay and they will increase in number and influence. Several SDGs will use indicators produced from official sources which may not use standardised methodologies and data sources. The different approaches may lead to discrepancies in the results which will have to be analysed and explained.

57. This work package considers the issues involved in integrating alternate data sources into the validation processes used for producing official statistics. Issues include assessing origin and quality of the source, trustworthiness and commercial or other interests of the parties exploiting them; designing processes and modelling techniques which are sustainable and formalised (as ad hoc adjustments to the statistics would be difficult to defend); and, educating users on proper use and interpretation of information (both the general public and more specific user groups).

58. In 2017 it is planned to expand the guidelines developed in 2016, including quality indicators on the quality of the validation process. The findings should be communicated to the HLG-MOS project developing quality indicators for GSBPM.

1. Brief description of the experiments

59. With respect to using data integration for the purpose of validating official statistics, two experiments were conducted:

- A comparative analysis of income data from New Zealand Income Survey with administrative data (New Zealand)
- Linking the Statistical Register of Employment and the Labour Force Survey (Slovenia)

60. This included the following activities:

- Identify issues related to systematically using data from other sources in the validation of official statistics
- Recommend potential approaches and modelling techniques

61. Work in 2016 focused on identifying different applications and methods for validating official statistics. Issues identified with use of administrative data to validate official statistics and lessons learnt from experiments as well as other validation projects carried out within organisation of contributing members are documented to provide initial guidelines in the use of administrative data to validate official statistics as well as recommend approaches and modelling techniques to resolve issues identified.

62. For 2017, it was proposed to continue the work to:

- Investigating the relevance of the ESSNET Rules Repository and ESSVIP validation definitions handbook as tools for validation
- Testing recommended approaches and modelling techniques
- Describing different applications of using data integration for validating statistics (e.g. validating existing statistics, replacing sources, design of new statistics, improving the design of existing statistics, challenging results from alternative sources/suppliers of statistics)
- Develop initial guidelines in the use of administrative data to validate official statistics. Include a description of an initial set of minimum requirements (these could include minimum metadata, process steps, methods) to initiate validation process.
- Determine what validation methods exist or could be created. Include software available to carry out validation methods.
- Document issues identified with use of administrative data to validate official statistics
- Recommend approaches and modelling techniques needed to resolve issues identified in the use of administrative data to validate official statistics
- Test recommended approaches and modelling techniques
- Document experience and lessons learnt
- Expand the guidelines in the use of administrative data to validate official statistics, if required. Include quality measures or indicators that are essential to report on the quality of the validation process. Communicate findings with the HLG-MOS project developing quality indicators for the GSBPM
- Develop training materials to carry out validation process using data integration

2. Related work in other projects/organisations

63. Related projects are carried out under ESSnet and ESS.VIP as follows:

64. ESS.VIP ADMIN Project: this project aims to find ways to optimise the use and accessibility of administrative data sources in the production of official statistics while guaranteeing the quality and comparability of these statistics. Details of work on this project are available from https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en.

65. ESSnet project on Data Integration: this completed project focused on the methodologies and methodological issues of micro data integration. Details of work on this project are available from http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en.

66. ESSnet project Integration of Survey and Administrative Data: the project aimed at developing the knowledge and expertise of participating NSOs in the use of integrated survey and administrative Data in the production of official statistics. Details of work on this project are available from http://ec.europa.eu/eurostat/cros/content/isad-finished_en.

67. ESSnet project on macro-integration: this project discusses various methods of integrating data sources at aggregated or macro level. Results of this project are available from https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

V. Opportunities, issues and challenges

A. Opportunities and potential benefits

68. Integrating data from multiple data sources allows NSOs to expand their use of external data sources in the production of official statistics and offers many **opportunities and potential benefits**. These are strong motivators for statistical offices to improve capacity in this area. Data integration can:

- provide more timely and more detailed statistics
- allow to create new official statistics or enhance existing official statistics with components from the external data source
- meet new and unmet data needs
- lower response burden
- overcome the effects of reduced response rates or replace existing data sources
- improve quality and address issues of bias in surveys

69. Data integration also allows improving the quality of the traditional statistical sources by finding the sources of error and determining methodological and operational issues that impact on quality. New data sources can provide a better sample or even full coverage. They can be cheaper than survey data and potentially of higher quality. Online data has the advantage of being fast and of high frequency.

70. In New Zealand, for example, the advancement of data integration skills has led to the creation of Statistics NZ's Integrated Data Infrastructure (IDI) bringing together linked datasets from a range of government agencies (including Statistics NZ's own data collections). The IDI is a large research database containing microdata about people and households and is continually growing. It has paved the way to answer complex research questions to improve outcomes for New Zealanders.

71. A special case of data integration **is integrating administrative and survey data.** This can serve different purposes: supplement sample surveys for a part of the population, for a set of variables, for estimation or for the data validation and editing process. In some cases, a sample survey can be replaced with data based entirely on administrative sources. Administrative data can also be a source for establishing and maintaining statistical registers, which are further used in implementing surveys.

72. Sample surveys are in general more flexible than administrative sources as they are designed to meet a precise purpose. Administrative sources on the other hand usually offer better coverage of target populations and in general have high response rates. It is cost-effective and cheaper to acquire data from existing administrative sources than to conduct a sample survey, and there is no additional respondent burden. As administrative data covers whole populations, local area data can be produced to a level of detail that is not permitted by sample surveys. This is also of advantage in implementing local policies. However, there are also challenges related to administrative source which are described in the next section.

73. **Integrating statistical data** and data form other sources **with geospatial information** adds value to both the statistical and spatial data. Other benefits of integrating with geospatial data are:

- improving quality of geospatial and statistical data

- enhanced collaboration between mapping agencies and statistical offices, improving maintenance and timeliness of data sets
- better interoperability of data sets, easier methods for linking data sets
- additional possibilities for (spatial) analyses and for presenting data
- new kinds of services and data to meet user needs
- flexibility for external users
- extended uses: for policy and decision makers, especially for regional policy makers; for scientific purposes; for environmental protection, etc.

B. Issues and challenges

74. Data integration comes with many challenges. The term data integration can be widely interpreted: there are many different types of data and different sources that can be integrated. Therefore different approaches are needed.

1. Legal and institutional issues

75. The first and most important legal issue is the **legal basis** for the use of administrative and other external sources for statistical purposes. Data integration projects carried out by NSOs is subject to legislation, codes of practice, protocols and policies, some of which are stipulated in their Statistics Act. The use of already existing administrative sources for official statistics may be included in national legislation. If it is not, there is a need to establish such basis.

76. For using other sources of administrative data there is often no legal basis. The fifth Fundamental Principle of Official Statistics says that “Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.” This gives NSOs the right to use data from different sources but it does not oblige the owners of other sources to share their data with the statistical offices.

77. Legislation with respect to **confidentiality** and **public perception** are of specific concern when integrating low granular administrative data, big data or spatial data.

78. It is important to manage the public perception and **communication** related to data integration to ensure continued trust in official statistics. Promoting understanding and support of data users and the general public of the data integration work being done within official statistics is an important issue. To assure public acceptance, the legislation related to data protection, e.g. Personal Data Protection Act should be followed. It determines the rules on processing personal data in a way that the legal rights of the individuals concerning privacy and integrity of individual's data are not violated.

79. **Collaboration with administrative data providers** is needed to ensure that the coverage, data quality, classifications, etc., used in administrative sources are in line with the statistical requirements. Collaboration of NSOs with administrative authorities in the preparation of legal documents establishing and maintaining an administrative source is a good solution to overcome this problem. The approval of NSOs in passing legislation on administrative records may be stated in a Statistical Act. Signed cooperation agreements may be needed to divide the tasks between the parties of the agreement, to define the rules and conditions of transferring data such as timeliness, technical implementation and metadata.

80. Data integration leads to opportunities to build **partnerships** for collaboration and sharing experience with statistical institutions, data producers, academia, research institutions, private sector, government institutions, commercial organisations, including providers of ICT infrastructure. These partnerships come

with all the challenges related to establishing strategic partnerships (which were considered at the CES seminar in April 2016). In some cases, the same data are used by several institutions, so continuous collaboration in institutional methodological groups is recommended to develop a system that is satisfactory for administrative and statistical purposes.

81. Moving data integration projects from research projects to repeatable, reliable production of statistics has its challenges, including **governance** of data integration projects. A number of countries have developed strong governance frameworks for managing data integration projects. For example, Australia has a Statistical Data Integration Framework ("National Statistical Service § Data Integration Need to know," 2014) which has been endorsed and is used across the Australian government. Other countries have similar frameworks.

2. Managerial issues, including resources

82. A general challenge with data integration is to have the right resources: the budget, human resources and IT infrastructure to provide the necessary **expertise, knowledge and technology**. Budget restrictions might restrain the ability to obtain the necessary resources.

83. **Human resources** include subject-matter statisticians, methodologists and IT experts. A mixture of a wide array of different **skills** is needed, both for working with data and for obtaining, negotiating and communicating the use of external data sources. These can include leadership, negotiations, relationship building, legal, data protection and communication skills, etc. Staff working on data integration should be aware of the various policies and legislative provisions that affect data integration, including the legislation that protects individual data.

84. The substantive skills include expertise in data content, statistical process, understanding of quality measures, etc. Having human resources with the right IT, software and database skills is crucial. Specific data integration projects, such as using big data may require knowledge of web-scraping tools, and IT skills for managing and processing big data. Geospatial expertise and knowledge of the relevant existing processes are necessary when using geospatial data sources.

85. Data integration is facilitated by the rapid development in the IT area, i.e. hardware equipment as well as a wide range of software tools. The **IT infrastructure** needed for integrating data covers servers, software and tools for developing databases where microdata and metadata are stored, and data processing and dissemination tools. Secure and efficient file transfer mechanisms are needed. Specific IT resources for storing and processing large volumes are requested for using big data. Geospatial data needs further infrastructural resources as data remains disaggregated beyond the initial processing on to dissemination. These services may be outsourced, in which case the requested funds to outsource should be available.

86. Data integration brings along **risks** that have to be managed. Official statistics have to produce stable output while the inputs for data integration are often unstable. As there is a chance on interrupted data sources, contingency plans should be in place in case the data source becomes unavailable.

87. To mitigate the risks, organisational risk management frameworks and Guidelines on Risk Management Practices in Statistical Organizations (currently being prepared by Istat) should be used. Some data sources may require specific risk assessment approaches.

88. Data integration may have high setup **costs**. Maintaining an ongoing data integration project also involves costs. Therefore, it can be recommended at first to focus on using limited resources to produce tangible and useful outputs.

89. External **data sources** are not necessarily available for free. Costs are also incurred to assess the quality of external data sources. Administrative data are usually cheaper than sample surveys as they are already being collected for administrative purposes, but they would still require some budget. Likewise, for using big data, time and the accompanying budget are needed for preparing of and experimenting with data.

90. The different potential sources for data integration may require using different methods. Therefore each data source should be tested/explored before embarking on an integration project. All the related costs need to be determined and assessed before proceeding with any data integration project.

91. Another challenge is **the resistance to changing any part of an ongoing production process** that will involve the integration of an external data source especially when current approaches are widely accepted and well-grounded expertise has been established.

92. Using **standard processes** which are common for different types of data integration would greatly facilitate data integration. Many data integration projects could follow a similar series of steps, such as:

- identifying needs and clarifying business requirements
- researching related work done in other organisations
- identifying partners and collaborators
- selecting potential data sources
- identifying methodological and quality considerations
- analysing whether and how the potential data sources can be used (as direct sources or for data validation)
- making a business case (including costs, benefits, risks, etc.)
- obtaining the data
- obtaining required tools, skills, resources
- experimenting
- assessing results
- refining methods and approach (as required)
- developing into a repeatable production solution

93. There are also a number of tasks for which it would be useful to have standard processes, such as:

- editing inconsistencies in linked records
- imputing values of a variable in linked records of an integrated dataset
- determining weights to adjust for missed links in an integrated dataset
- carrying out validation methods

3. Methodology, concepts and definitions

94. Methodology of data integration depends largely on the data sources used and there are **no common recommended methods** for integrating all types of data. It is planned to form a catalogue of commonly used and new methods for the different types of data integration under the HLG-MOS data integration project. Ideally, the methods should be described in a form that is compatible with the Common Statistical Production Architecture.

95. Most of the data sources to be integrated are external to the NSO. The NSO had no control over the collection of the data and differences in **concepts, classifications, populations and collection units** are expected.

96. When integrating data, the statistical concepts should be aligned with the **concepts** in new data sources. Sometimes new concepts need to be developed. An NSO need to understand the design decisions so they can determine how to turn external data into statistical information. These differences affect the usability of the

external data source in the production of statistics specifically with regard to: the coverage of population, the validity of the target concepts, the availability and accuracy of descriptive metadata, sampling error, bias, legal basis for data, data collection methodology/questionnaire design, response burden, confidentiality of the resulting output, and different consequences for different types of data provided. These differences need to be clearly explained and documented, and stored to ensure reuse and improvement of assessments.

97. **Users** of statistical information should be well informed about the definition of the concepts and populations used in all data sources used in a statistical output. A common understanding of statistical concepts between an NSO and users of the statistical information should also be ensured.

98. For example, when **integrating administrative and survey data**, **t**he difference in concepts might lead to **coverage** problems as well as **bias** problems. In some cases, such as business statistics, **units** used in administrative data do not necessarily correspond to the definition of the required statistical units. Some modelling should be done to convert the administrative units into statistical units. It is likely that there will also be differences in the definitions of variables. It is important to have a thorough knowledge of the impact of these differences. Sometimes it is possible to influence the administrative definition by co-operating with the responsible authority.

99. In cases of different **classifications**, the usual step is to use correspondence tables and conversion tools based on additional variables that may be available for converting into more correct classification code. However, even the same classifications may result in different data, especially when classifications are complex or the rules of a classification are difficult to apply. In administrative sources, there would often be respondent coding, while a sample survey may have open questions and coding is often done by experts.

100. Cooperation between NSO and the administrative authority is a good way to solve a part of the classification problem. NSO can provide experience and may be the one responsible for maintaining the classification. Another issue is whether to use directly translated international classifications or national classifications. It depends on what national data are needed. The first option is usually harder to implement in case of changes and revisions compared to having national classifications. To change a classification in an administrative source is a demanding task since there can be many data providers that need to become familiar with the changes.

101. When integrating geospatial with other data sources, the granularity or aggregation level of the data sets might not coincide and coverage might be different.

102. Problems to overcome are also the **missing data and errors**. In statistical surveys, the missing data are due to unit or variable non-response, but in administrative sources the causes can be different. It is important to identify if errors and missing data are systematic, and apply appropriate validation and editing rules.

103. A common issue with linked datasets is inconsistencies in the records linked. As the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records increase as the number of variables increase. However, this may also increase the amount of editing required.

104. Sources of potential bias should be identified with regard to integrated datasets. Coverage and conceptual issues may only apply for some groups of a population so care should be taken in generalising results. Some variables can affect the quality of linking and may be a source for potential bias in carrying out analysis

on resulting datasets. Investigations on linkage rates across different subpopulations may be required.

105. **Standardization of identifiers** (or other information by which it is possible to link statistical information) across different sources is one of the most important aspects in the integration of administrative data. If there is no such system, it is much more difficult to link different sources and data linking and matching methods must be applied.

106. Methods to better estimate linkage errors are required to determine appropriate models to account for linkage errors. Linkage errors contribute to potential coverage errors in the resulting target population. Care should also be undertaken when creating statistical units from integrated datasets where one dataset is external since the unit may be defined differently in the external dataset.

107. Extreme care should be taken in fore- and back-casting linked data especially for longitudinal data. A person may link in one quarter but not in another due to data quality reasons (or may link to a different record). The use of linked datasets even for validation purposes may result in a break in the data series that needs to be managed.

108. Data sourced externally may suffer from measurement errors, e.g., validity error, and these errors propagate when the data is integrated with other data sources to produce a statistical output. Hence, target concepts used in a dataset sourced externally should be well understood before being used in the production of official statistics.

4. Quality

109. **Quality** of both the data sources and the statistics produced have to be measured, managed and published. The following dimensions of quality need to be assessed: accuracy, relevance, consistency, accessibility, comparability and timeliness. It is important to have detailed descriptive **metadata** to assist in the assessment of the quality of the data sources. Data integration makes metadata more challenging. The minimum and ideal metadata required should be identified.

110. A **quality framework** would be needed for integrated data to understand the "uncertainty" or the reliability of the estimates, how comparable the indicators are over time and with similar indicators produced from traditional sources. A quality framework aimed at determining the optimal design of combining data source(s) that can minimise the cumulative effect of potential errors on a statistical output is essential. The draft quality framework for Big Data produced under the 2014 HLG-MOS Big Data project and the traditional quality frameworks can be used as a basis to develop a first draft of a quality framework for statistics produced from integrated data sources.

111. **Timeliness and differences in reference period** are challenges in integrating data sources. Administrative or spatial data may not be available in time or may not coincide with the statistical reference period. It can be resolved by analysing the impact and if necessary adjusting it by models.

5. International cooperation

112. Using the collective experience of the official statistics community is important to avoid duplication of effort across countries and organisations and **learn from each other's experience.** Sharing experiences among official statistics organisations will help to develop and use common approaches in pursuit of more efficient and effective inclusion of data integration into the statistical production processes.

113. **Forum for international collaboration** in solving common problems would be welcome. Coordination with international groups is needed where relevant.

114. There is potential to bring some of the data providers and a group of NSOs together to explore mutual benefits and potentially develop agreements for data supply. When using IT resources outside the office, good partnerships with developers and administrators of such systems is needed.

115. As guiding tools, **standards** developed under the umbrella of the HLG-MOS, such as GSBPM, GSIM and GAMSO should be used and where possible, CSPA compliant services should be developed or proposed. We also need to identify other common standards and frameworks used in the wider environment that are relevant to data integration tasks.

116. The lack of well-established standards for integration of different data sources provides the opportunity to propose new standards. Concordance between statistical and other standards is needed. Therefore standards used by source industries have to be considered (e.g. education, travel, banking).

117. Developing common approaches for particular areas of statistics would be useful. The HLG-MOS sandbox and sample data sets can be used for this purpose. NSOs from different countries can work together to develop a common approach for obtaining data from multinational companies.

VI. Conclusions and recommendations

118. **Data integration provides the potential to produce more timely, more disaggregated statistics at higher frequencies than traditional approaches alone.** Data integration activities will therefore only increase. With ever more data sources becoming available and increased capacities of IT and data infrastructure, the need for integrating different sources will grow.

119. For some types of data integration experiences exist at certain national offices. In many cases experiences are more limited and no office has expertise with all types of data integration. **No or limited guidelines exist for data integration and a comprehensive overview of experiences is missing.** Collaboration and exchange of experience with integration can be valuable for the international statistical community.

120. Continuation of the HLG-MOS Data Integration project and **increased sharing of experiences and intensification of collaboration between national statistical offices and other agencies producing official statistics as well as data providers is strongly recommended.** Sharing and collaboration will lead to cost savings and synergy to advance developments at national and international levels.

121. The various data integration experiments lead to many recommendations. Here is a summary:

- get high level management support and let management of statistical offices know about the project and its goals and results
- keep your objectives clear and clarify objectives and importance of the project, specify the characteristics of the problem to be resolved
- good collaboration with partners/data providers: be clear about the data requirements, take into account the common goal of institutions (collecting data only once, reducing unnecessary expenses), set up agreements
- collaborate with users of the data
- consult with other experts, find out more on possible methods and solutions in a practical sense; exchange personnel to gain experience learn from good practices
- consider international recommendations and standards for statistics
- share results

- start the work on a sample, look at the data before commencing, pertain patience and persistence, have in mind a wide range of solutions
- measure data quality

122. With respect to the last recommendation on data quality frameworks, more extensive recommendations have already been produced by the project. These are available on the HLG-MOS wiki at http://www1.unece.org/stat/platform/x/axSzBw.

123. **Bureau members are invited to:**

- **comment on the three main conclusions in paras 119-121**
- **nominate their experts to join the 2017 HLG-MOS Data Integration Project**, particularly in the production of the synthesis report. The tasks include drafting and finalising guidelines and recommendations, and making them available as an on-line *Guide to Data Integration* for the statistical community
- **advise on how to further develop the work on data integration**, including the maintenance and promotion of the online guide after the current project ends in December 2017.

* * * * *