A Guide to Data Integration for Official Statistics

(Version 2.0)

Table of Contents

Preface

The High Level Group for the Modernisation of Official Statistics (HLG MOS) sponsored projects on Data Integration in 2016 and 2017. The following countries and organisations participated in the projects: Brazil, Canada, Colombia, Hungary, Italy, Mexico, Netherlands, New Zealand, Poland, Serbia, Slovenia, United Kingdom, and Eurostat.

The guide uses information gained from national and international statistical organisations and from other data integration work completed or in progress, to provide practical advice and information to advance data integration activities by statistical organisations.

The guide is relevant to managers, statisticians, methodologists, ICT professionals and other staff in statistical organisations who are using, or planning to use, data integration in the production of official statistics.

It provides information about issues that statistical organisations have or should consider in work on data integration. This information can be used to:

- Speed up development of data integration strategies and practices
- Support consideration of factors that may not have been otherwise considered
- Encourage joint approaches or reduce redevelopment of work already done within the official statistics community
- Find examples of similar work done in other statistical organisations for specific statistical domains or using specific technologies, methods or approaches.

I. Introduction

1. Increasingly, new data sources are becoming available to statistical organisations. This comes at a time when modern technologies are available to support data integration. Data integration provides the potential to produce timelier, more disaggregated statistics at higher frequencies than traditional approaches alone.

2. It can be used to provide new official statistics, address new or unmet data needs, lower response burden, overcome the effects of reducing response rates, and address quality and bias issues in surveys.

3. Statistical organisations are challenged to integrate diverse sets of inconsistent data and to produce stable outputs with sometimes unstable, ever-changing inputs. Instead of trying to produce the best possible statistics from a single survey, it is necessary to try to find the best combination of sources to deliver the indicator/statistics that best satisfy the users' needs.

4. There are some potential challenges related to data integration including:

- New skills, new methods and new information technology approaches
- Designing new concepts or aligning existing statistical concepts to the concepts in new data sources
- Measuring, managing and publishing the quality of both the data sources and the statistics produced
- Governance for data integration projects
- Managing public perception and communication
- Avoiding duplication of effort across countries and organisations and using the collective experience of the official statistics community.

5. A survey conducted in 2017 by the Data Integration Project[1] found that public acceptance and trust issues caused more significant/moderate barriers to data integration compared to issues such as methodologies, skills and budget.

6. This guide provides information on planning for data integration activities, issues related to data and the data integration methods and tools.

II. What is data integration?

A. Definition

7. In order to provide guidance for statisticians on how data integration activities fit into the statistical business process, it is important to define the term "data integration".

8. According to the Generic Statistical Business Process Model (GSBPM[2]), data integration

[1] https://statswiki.unece.org/display/DI/Results
[2] https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

is an activity in the statistical business process when data from one or more sources are integrated. Although the GSBPM defines data integration under Process, it is important to remember that the GSBPM is not a linear model. Data integration is possible in the development and production and dissemination of official statistics whenever a combined, integrated dataset is produced.

9. Data sources could be a mixture of various data sources. In official statistics, these sources are usually primary data sources (more "traditional" sources such as statistical surveys) or secondary data sources (typically administrative datasets, big data or any other non-traditional source of information for official statistics). The result of the data integration activity is always an integrated dataset.

10. This document defines data integration as **the activity when at least two different sources of data are combined into a dataset**. This dataset can be one that already exists in the statistical system or ones that are external sources (e.g. administrative dataset acquired from an owner of administrative registers or web-scraped information from a publicly available website).

11. Some examples of data integration include:

- an integrated dataset that serves as an input to produce official statistics
- a statistical model developed and produced using different sources to produce model-based information
- a dataset integrated for the purposes of micro-validation when some rules are defined to check the validity of the data in one dataset compared to another one
- missing values imputed in a dataset using another dataset as the source for imputation
- datasets combined to produce a sampling frame for a survey
- data from several subject-matter domains combined into one dataset that is the basis for the production of statistics (example: national accounts)
- datasets from different subject-matter domains compared to check the quality and the validity of information produced (macro-validation)
- input from several sources integrated into one dataset to provide microdata files for the researchers for scientific purposes
- different sources used to apply proper statistical disclosure control methods on a microdata set.

B. Types of data integration

12. There are many possible types of data integration. Five common types of integration are: administrative sources with survey and other traditional data; new data sources (such as big data) with traditional data sources; geospatial data with statistical information; micro level data with data at the macro level; and validating data from official sources with data from other sources. More information about these can be found in Annex 1.

13. 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.

14. The survey conducted in 2017 asked statistical organisations about their data integration experiences and practices. The results showed that the four most comment types of data being

integrated with other data sets were survey data, census data, commercial transactions and data from public administrations.

15. Data integration techniques can be applied for several reasons in the statistical business process. The 2017 survey found that data is commonly used to supplement survey data (e.g. for part of a population, for a set of variables), validate data, maintain registers and edit/impute data. It is also used as a source for sample frames.

16. The results of the data integration survey show that data integration for the ongoing production of statistics is more commonly used in some statistical domains (for example Business Statistics and Economic Accounts) than other domains. There is a high level of experimentation and research in some of the other domains (for example, Education and Indicators related to the Millennium or Sustainable Development Goals).

17. There are a number of groups which are working on developing strategies for global data agreements and methods in different statistical domains. Two examples are the Task Force on Data Integration for Measuring Migration [3], the Ottawa Group[4] and the Group of Experts on Consumer Price Indices [5] which are working on issues such as the use of big data and global data agreements for consumer price indices.

III. Planning for Data Integration

18. A number of statistical organisations have developed or are developing international, national and/or organisation wide strategies for data integration.

19. Some of the most common activities when planning for data integration include the use of co-operation agreements for transferring data, preparation of legal documents for establishing and/or maintaining use of the data, and developing long-term partnerships (formal or informal) which consist of two or more institutions using the same data.

A. Access to data

20. A key requirement is to have access to the desired data. In most countries, data can be accessed freely by the statistical organisations. In some countries and cases, data are only available with payment and in some cases the data exists but cannot be accessed.

21. A legal basis is often important to access the data for statistical purposes. A sound approach is to ensure national legislation is aware of already existing administrative sources rather than recollecting data. The usage of administrative sources is often stated in a Statistical Act. It is needed to consider various statistical, methodological, legal and ethical issues.

22. Practical work to develop common approaches starts with data. Some types of data can

---

[3] https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2017/mtg1/2017_UNECE_Migration_WP_18_TFDataIntegration_DraftReport_ENG.pdf

[4] https://unstats.un.org/unsd/methodology/citygroups/ottawa.cshtml

[5] http://www.unece.org/index.php?id=41278#/

be used without much difficulty, like web-scraped data or social media sources, government-owned open data or public statistical outputs. In other cases, already available public use files or IPUMS data[6] may be made available. It is also possible to systematically anonymize real data from surveys, censuses, government registrations or privately held big data sources.

23. At the beginning of new collaborative data integration projects, use of experimental data sets makes this easier. The creation and documentation of a set of synthetic datasets allows countries to collaborate on developing common methods, removing issues of confidentiality and encouraging use of the same data formats. 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. There is potential to bring some of these data providers and a group of statistical organisations together to explore mutual benefits and potentially develop global or regional agreements for data supply.

24. An example where many countries have similar challenges which can be assisted with data integration is improving the quality of the Consumer Price Index (CPI) in terms of coverage and real-time quantity. Several data providers operate in many countries and there is an opportunity to develop a common approach that can be used in multiple countries[7] [8]. Some organisations have significant data holdings (for example IRi [9] with 5 years of supermarket scanner data).

25. There are other issues to be considered in securing access to data. These include:

- Access to application programming interfaces (API) can be restricted under terms and conditions
- Some countries may not have access to the same data sources or there may be different formats, various levels of detail or aggregation across countries
- Data may no longer be representative of current situation
- Commercial companies may not be interested in partnering with statistical organisations unless a compelling case is made
- There is potential for interrupted supply
- If the statistical organisation can't get data, they could may still be able to generate and provide meta-information about the data
- Use of a common area for the lodgement and storage of data (not necessarily in statistical organisations) is helpful
- There is a need for secure and efficient file transfer mechanisms for data used in production processes.

B. **Partnerships**

26. The importance of establishing and maintaining effective partnerships for data integration should be recognised. The types of partnerships include:

---

[6] www.ipums.org
[7] www.gfk.com
[8] www.pricestats.com
[9] www.iriworldwide.com

- collaboration and sharing experience, approaches and standards with other official statistical organisations
- partnerships with data providers
- cooperation among institutions within countries (e.g. tax offices, employment office) – together deciding the methodology, concepts and classifications
- public/private partnerships with technology organisations, research initiatives and academia.

27. There are many factors which encourage effective partnership. These include:

- establishing personal and friendly connection between organisations - relationship management
- providing feedback on the data regarding its usefulness for official statistics needs
- promoting the goals of the data integration project to the providers and jointly clarifying mutual benefits
- understanding and managing barriers such as costs, capacity and risks
- establishing formal agreements.

C. Skills

28. Some of the essential skills needed for integrating data are:

- leadership and negotiation skills are useful for participating in policy development and in discussions with data providers
- legal skills relate to the legal basis for obtaining data, data protection and a co-operation agreement between the administrative authority and the statistical organisation
- subject-matter statistician skills cover expertise in knowing data content, understanding and analysing data, knowing the statistical process and dissemination methods
- methodological skills relate to all statistical processes such as sampling frame preparation and selection of observation units, data linkage and matching, weighting, time series analysis and seasonal adjustment, data protection, etc.
- programming, software and database skills are needed for construction of microdata databases and for establishing and maintaining generic and non-generic process programs (e.g. for data editing and imputations, validation, aggregation and tabulation, micro and macro data analysis, data protection).

29. The Survey on Data Integration asked organisations about the level of skills and interest in obtaining or providing skills development related to data integration. Interestingly, organisations indicated the areas where training was most required with methods and quality frameworks. Information on these topics can be found in Sections V and VI.

IV. Data Considerations

30. There are a number of issues to be discussed when considering the data sets to be integrated. These include the conceptual alignment of the datasets, identifiers and privacy concerns. The following sections outline these issues.

A. Concepts

31. Most of data sources to be integrated are external to the statistical organisation. The statistical organisation does not always have control over the definition of the concepts and populations used in the collection of the data.

32. The data to be integrated needs to correspond to statistical concepts. Administrative and data form other non-traditional sources are primarily collected for non-statistical purposes. Therefore, there are often differences from what is required in statistics such as differences in concepts, coverage, units, definition of variables. There is an important need for detailed descriptive metadata to assist in the assessment of the quality of the data sources. The following dimensions of quality need to be assessed: accuracy, relevance, consistency, accessibility, comparability and timeliness.

33. Using someone else's data means a statistical organisation cannot control any of the decisions on measurements and populations undertaken by an external data source provider. A statistical organisation need to understand the design decisions, so they can determine what to do to turn external data into the statistical information they want.

34. These differences affect the usability of the external data source in the production of a statistical product 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, by product data versus survey question, confidentiality of the resulting output, and different consequences for different types of data provided. These differences need to be clearly explained, documented, and stored to ensure reuse and improvement of assessments. Good quality variables closely related to each other in different datasets would be ideal to use for linking.

35. Collaboration with the data provider is one way to lower the risks. This is especially applicable in the case of administrative records which are collected for the purpose of implementing various non-statistical programs concerning legal requirements such as taxation, housing, pensions, social benefits, trade in goods, etc. Both the provider and the statistical organisation have an interest in quality, but the relevant quality aspects and priorities can be different for the production of statistical data. Statisticians may have to make compromises concerning coverage, data quality, classifications, etc., in administrative sources.

36. Collaboration of the statistical organisations 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 the statistical organisations in passing legislation on administrative records may be stated in a Statistical Act.

37. Control of the methods by which the administrative data are collected and processed rests

with the administrative agency. They are specialized in formulating transparent rules and procedures. The statistical organisations have experience in data collection, classifications and data validation. 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. When acquiring data, cooperation agreements are signed 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.

B. Identifiers

38. Another requirement for data integration is connectivity. This is easiest with a unified identification system across different sources. In many countries, unified identity systems exist for persons, businesses, farmers and addresses (or geo codes). Often the identity numbers are anonymized and translated into statistical identity numbers for privacy protection in the statistical production.

39. If there is no unified system in the country, it is much more difficult to link different sources. If the sources contain unique identifiers, the integration is directly achieved via these identifiers; otherwise, it is necessary to define and prepare a procedure for pooling records by different parameters (indirect integration).

C. Privacy and Confidentiality

40. Integrating and holding/storing more data sources increases disclosure risks and therefore needs to be managed carefully. To assure public acceptance, privacy and confidentiality rules must be also clear. Privacy refers to the freedom from intrusion into one's personal information. Confidentiality refers to personal information shared with others. Confidentiality means that the information can be assessed only by authorized individuals.

41. A Personal Data Protection Act 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. The ability to integrate data sources in national statistics also depends on the trust of observation units: persons, households, enterprises, agricultural holdings and other organizations. This means that respondents and administrative sources will share their data if they are convinced that the confidentiality of the data and identity is ensured and that the shared data will only be used for statistical purposes.

42. The statistical organisation needs to provide information and explanations of the applied procedures. The protection (safeguarding) of confidentiality also aims to ensure that the disseminated data do not allow direct identification (via direct identifiers) or indirect identification (by any other means). This confidentiality must be protected under legislation. The mission of national statistics is to transmit and release statistical results to the widest extent possible while minimizing the risk of the disclosure of information on units. To this end, appropriate statistical disclosure methods are needed to ensure compliance with the legislation.

V. Quality

43. Quality assessment is recognized as an important issue in statistical production. Many organisations and international and national initiatives have considered various aspects of quality and some of them explicitly consider processes related to data integration. The following dimensions of quality need to be assessed: accuracy, relevance, consistency, accessibility, comparability and timeliness[10].

44. A useful reference for the quality measures in data integration steps is the Quality Indicators for the Generic Statistical Business Process Model (GSBPM)[11]. In this work, indicators to evaluate the quality of standard linkage procedures are proposed.

45. When integrating survey data and administrative data, the ESSnet Komuso (Quality of multisource statistics)[12] provides useful documents to:

- Take stock of the existing knowledge on quality assessment and reporting and review it critically in order to produce recommendations on the most suitable approaches;
- Develop new indicators for the quality of the output based on multiple sources;
- Produce a methodological framework for reporting on the quality of output;
- Produce indicators relating to the quality of frames themselves and the data whose production is supported by frames;
- Produce a methodological framework for assessing the quality of the frames used in social statistics; draft a proposal for minimum quality requirements for sampling frames for EU social statistics;
- Produce recommendations on updating the ESS Standard and the ESS Handbook for Quality Reports.

46. The outputs of the "Methodologies for an integrated use of administrative data in the statistical process" (MIAD[13]) project provide a generic framework to assess the quality of the administrative data at the input stage, quality indicators for the discovery phase and acquisition phase, and a guide to reporting the usability of an administrative data source.

47. At the national level, statistical organisations recognize the necessity of developing a framework for assessing quality in the usage and integration of different data sources. The following resources are recommended:

- Statistics Canada Quality Guidelines (2009)[14],
- Checklist for the Quality evaluation of Administrative Data Sources of Statistics

---

[10] "Guide to reporting the usability of an administrative data source" deliverable B3 of the MIAD project https://ec.europa.eu/eurostat/cros/system/files/Guide%20to%20report%20the%20usability%20of%20an%20ADS.pdf

[11] https://statswiki.unece.org/display/GSBPM/Quality+Indicators+Home

[12] https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en

[13] https://ec.europa.eu/eurostat/cros/content/miad-methodologies-integrated-use-administrative-data-statistical-process_en

[14]http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm.

Netherlands (2009)[15]

- Quality Guidelines for Statistical Processes of Istat (2016)[16]
- A Quality Framework and Case Studies from Statistics NZ [17]
- UK Statistics Authority Quality Assurance of Administrative Data[18]

48. The quality assessment framework, including the quality indicators, is described in Guide to reporting on admin data quality[19] is helpful in carrying out validation studies. The quality framework is based on Li-Chun Zhang's two-phase life-cycle method model for integrated statistical microdata [20](Figure 1) which expands the total survey error paradigm to include administrative data.

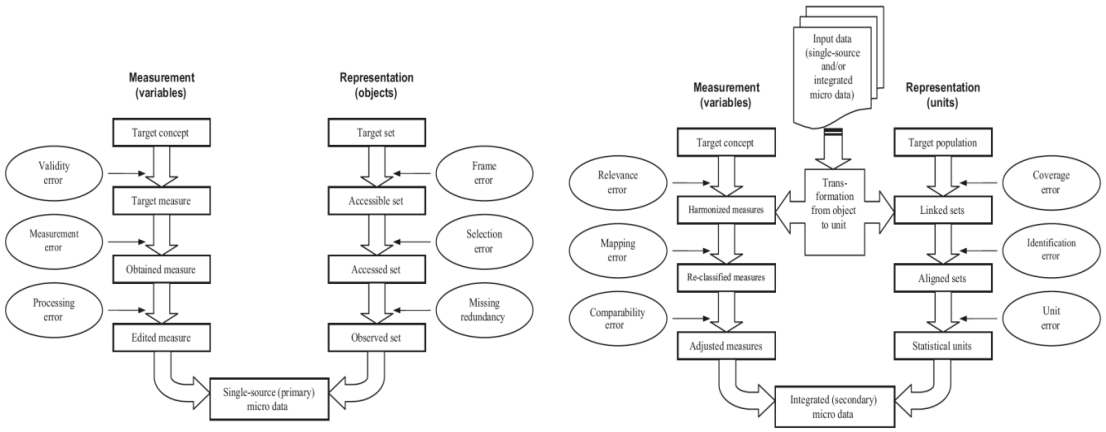

Figure 1. Zhang's two-phase life-cycle method model for integrated statistical microdata

49. The framework enables understanding of the error sources from the individual data sources including those arising from the integrated datasets. Zhang's two-phase life-cycle model assists in determining the associated methodological and operational issues that may impact on quality resulting from producing statistical information from linked administrative data sources.

50. Phase 1 assesses the quality of an input data source that is intended to be used in the production of a statistical product. A statistical organisation needs to understand the design decisions undertaken by the producers of the source to determine methods to turn the data into the statistical information required by the statistical organisation. Quality of the input data source is assessed against the purpose for which it was collected. For a survey dataset, this purpose is defined for a statistical target concept and target population. For an external data source, the entries or 'objects' in the dataset might be people or businesses, but they could also be transaction records, or other events of relevance to the collecting agency. At this stage, evaluation is entirely with reference to the dataset itself, and does not depend on what a statistical organisation intends to do with the data. Quality issues in the input data source will flow through into any use of the data in the production of a statistical product.

---

[15] Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Toth, J. (2009), Checklist for the Quality evaluation of Administrative Data Sources. Statistics Netherlands, The Hague/Heerlen

[16] http://www.istat.it/en/files/2013/04/Linee-Guida-v1.1-Versione-inglese.pdf

[17] http://dx.doi.org/10.1515/JOS-2017-0023

[18] https://www.statisticsauthority.gov.uk/gsspolicy/quality-assurance-of-administrative-data/

[19] http://archive.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx

[20] http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/abstract

51. Phase 2 categorises the difficulties arising from taking variables and objects from source datasets and using them to measure the statistical target concept and population a statistical organisation are interested in. In this phase, the statistical organisation considers what they want to do with the data, and determine how well the source datasets match what they would ideally be measuring.

52. The quality assessment involves 3 steps.

Step 1: Initial metadata collation: Basic information is collected about each of the source datasets used in the validation project. The information relates to the source agency, purpose of the data collection, populations, variables and timeliness of the data.

Step 2: Phase 1 evaluation: Errors occurring in phase 1 of the quality framework are determined and categorised for each source dataset. This involves detailed consideration of how the methods, purpose, known issues, and other aspects of the original data collection contribute to each of the specific error categories in the phase 1 flow chart in figure 1.

Step 3: Phase 2 evaluation: As for the previous step, errors arising in phase 2 of the quality framework are listed and examined in a similar way, taking into account the dataset(s) being integrated to produce the final output. These errors are considered with respect to the intended statistical target concepts and population. The effects of phase 1 errors on the creation of statistical units, or the particular details of the misalignment between concepts on different datasets, must be understood.

53. The Guide to Reporting the Quality of Administrative Data provides a metadata information template that encourages thinking about the key aspects of quality in an organised way. It is also a convenient way to record a standard set of information to compare different datasets. The basic information required are: name of data source agency, purpose of data collection, time period covered by the data, the population (target and actual) population of the dataset, the reporting units, a short description of key variables and the timing/delay information and method of collection.

54. For the integration of Big Data sources in the statistical production, several initiatives at international level have stated the potential of these new data sources, as well as the quality issues related to several aspects. Firstly, the change of paradigm imposed by the new sources, compared to the traditional sample surveys, moves attention from the well-studied sampling errors to the non-sampling errors, so the population coverage and the self-selectivity of the observations become the most recognized and investigated issues. A general framework for assessing the Quality of Big Data has been prepared by the HLG-MOS Big Data project[21].

VI. Methods and Tools

55. Data integration procedures differ regarding the types of data sources to be combined, the

[21] https://ec.europa.eu/eurostat/cros/system/files/Task%20Team%20Big%20Data%20Quality%20Framework_937_unblinded_v1.pdf

characteristics of data sets – such as the coverage and overlapping of data sets through the data sources, the micro- or macro level of data, the existence and usability of unique identifiers, and the purposes of combining data.

56. Data integration methods can be divided into two main groups: a) Record linkage methods and b) Statistical matching methods, that, however, can be further divided in several subgroups and categories from different perspectives.

57. There is growing literature available on these methods, on their sub processes, advantages and disadvantages, mathematical bases and tools. Indeed, there have been some international projects and training programs related to data integration – such as the CENEX-ISAD (European Centres and Networks of Excellence, Integration of Surveys and Administrative Data[22]), the ESSnet on Data Integration (ESSNET[23]), the CULT project (CULT[24]) or the European Statistical Training Program on Statistical Matching and Record Linkage in 2016 (ESTP[25]) that reviewed and contributed to the specialized literature on data integration with important reports, articles, lists of recommended bibliographies and other materials. It is strongly recommended the reader to check these projects, their work packages, reports and outputs at the web pages shown in the footnotes.

58. To avoid duplication, a brief overview of data integration methods, with some of their most notable features, and some of the tools that are used in official statistics to carry out data integration are given.

A. Record Linkage

59. *Record linkage* refers to the identification and combination of records corresponding to the same entities – persons, enterprises, dwellings, households, etc. – throughout two or more data sources. Record linkage methods can be further classified in two branches:

a) *Deterministic matching* (or exact matching) is when a formal decision rule – usually the coincidence (or mismatch) of the unique identifiers that correspond to the same units in two or more data sources – is applied to find out whether a pair of records is a match or not;
b) In the case of *Probabilistic matching* such strict decision rules are not applicable. Instead, complex probabilistic decision rules are established based on a set of key variables that are common in the data sets to be integrated to be able to accept or refuse matches on a probabilistic basis.

---

22 http://cenex-isad.istat.it/

23 http://ec.europa.eu/eurostat/portal/page/portal/essnet/data_integration

24 http://www1.unece.org/stat/platform/display/statnet/CULT+Project

25 https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2016%20ESTP%20PROGRAMME/36.%20Statistical%20matching%20and%20record%20linkage%2c%2019%20%e2%80%93%2021%20September%202016%20-%20Organiser_%20DEVSTAT

60. Due to the similarities of deterministic and probabilistic matching methods, i.e. that both are based on the matching of key variables, some of their features are common. First, the deterministic and probabilistic matching procedures both can lead to linkage errors as false matches (or false positives) interpreted as real ones or false "unmatches" (or false negatives) that is, real matches not recognized as such. Moreover, both consist of similar phases (For further details see for example the presentations of the ESTP course on Statistical Matching and Record Linkage, 2016, or the CENEX-ISAD WP1.):

1) Pre-processing:
   - Choice of the key variables,
   - Data cleaning and quality improvement,
   - Key variables in standard forms.
2) Linkage:
   - Match (same entity);
   - Unmatch (different entities);
   - Uncertain match (unable to decide – possible match).
3) Post-linkage (Manual review of unlinked records)
4) Data analysis.

61. *Deterministic matching* is considered as the ideal case of record linkage due to the existence of a unique identifier – social security number of persons, fiscal code of enterprises, geocodes of addresses, etc. – which assures an error-free, one-to-one matching of records with the same identifier, that is, that belong to the same entity. For this reason, there is considerably less literature on this method than on the others. However, some challenges can emerge during the application of this method. A possible difficulty is that unique identifiers could also be affected by errors occurred for instance during either the data collection, or the data capture processes. There could be missing values as well in some of the data sources. Identifying records in basic registers– in the "spines of integration" – could be a proper solution in order to obtain or check unique identifiers.

62. Contrarily, *Probabilistic matching* is a more complex approach. Instead of unique identifiers, softer key variables are used here such as the name, date of birth, address, or other variables describing the units of the target population. These are more prone to be affected by data collection or data capture errors, or they are often recorded in different formats making their comparison more complicated. In these cases, the pre-processing phase has a crucial role that could strongly affect the results of the record linkage exercise.

63. The complex mathematical bases of probabilistic record linkage and probabilistic decision rules go back to the ground-breaking works of Newcombe et al. (1959) and Fellegi - Sunter (1969) who formalised the theory of probabilistic matching based on the assumption of conditional independence. Even today, this method serves as the basis of record linkage applications. Other probabilistic record linkage techniques are that of Jaro's (1989), further developed by Winkler (1995), or the distance-based record linkage method as described by Pagliuca and Seri (1999).

Tools for record linkage

64. An up-dated critical review of methods and software for record linkage is in Tuoto *et al*. (2014)[26] where the proliferation of methodologies and tools in recent years is interpreted according to assigned criteria, namely flexibility of the tools with respect to the support to input/output formats, extensibility, maturity, supported language and coverage of functionalities related to identified sub-phases in which it is possible to organize a record linkage process to reduce its recognized complexity.

65. The CENEX-ISAD WP3 report and the CULT projects results offer a detailed discussion of software tools for data integration. These include:

- Automatch
- Febrl
- GRLS
- LinkageWiz
- RELAIS
- DataFlux
- Link King
- Trillium Software
- Link Plus
- RecordLinkage (R codes)
- FRIL
- Fundy
- QualityStage

B. Statistical Matching

66. *Statistical matching* (or synthetic matching) involves the integration of data sources with usually distinct samples from the same target population, in order to study and provide information on the relationship of variables not jointly observed in the data sets. The main difference from record linkage – as Leulescu and Agafit (2013) put it – is that "record linkage deals with identical units, while statistical matching deals with 'similar' units. In practice, matching procedures can be regarded as an imputation problem of the target variables from a donor to a recipient survey." The statistical matching situation is usually described with a recipient data source A containing X and Y variables and donor data source B with X and Z variables. That is the statistical matching itself is imputing Z variable in data source A using the common variable X.

[26] Tuoto, T., Gould, P., Seyb, A., Cibella, N., Scannapieco, N., Scanu, M. (2014) Data Linking: A Common Project for Official Statistics in Proceedings of CONFERENCE OF EUROPEAN STATISTICS STAKEHOLDERS Rome 24/25 November 2014

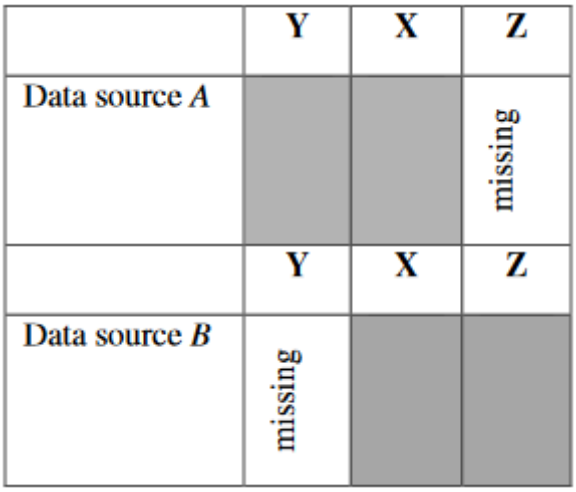

Figure 2. Statistical matching illustration by Eurostat (2014)

67. Statistical matching methods are categorized in the specialised literature from different angles:

a) The *Micro approach* aims at constructing a complete (containing all variables of interest) and synthetic (that is of not directly observed units) micro level data set.
b) The *Macro approach* seeks the integration of data sources in order to facilitate the estimation of the parameters of interest as the correlation or regression coefficients, and contingency tables of not jointly observed variables at the macro level

68. At another level, the Micro and Macro approaches can both be parametric or non-parametric and a mix of them can also be applied for the Micro approach:

c) The *Parametric approach* is based on the normality assumption of data. In this case a specified model is needed for the joint distribution of the variables that can lead to misspecification. (Usually maximum likelihood).
d) The *Non-parametric approach* is applied when data do not hold the normality assumption. This approach is more flexible than the parametric one when variables are of different types. (Usually hot-deck technics).
e) A *mix* of the parametric and the non-parametric approaches can be applied in the case of micro level matching: "first a parametric model is assumed, and its parameters are estimated then a synthetic data set is derived through a nonparametric micro approach. In this manner the advantages of both parametric and nonparametric approach are maintained: the model is parsimonious while nonparametric techniques offer protection against model misspecification" (D'Orazio, 2017).

69. Furthermore, approaches can be distinguished in accordance with the availability of information on the not jointly observed variables:

f) Approaches that assume the *conditional independence* of variables (originally all the micro-, macro-, parametric, non-parametric, and mixed methods were based on the conditional independence assumption).
g) Approaches where *auxiliary information* is available from a third data set in which variables are jointly observed.
h) In the case of *Uncertainty*, no assumptions are made, and no joint information is available on the variables, thus uncertainty analysis technics are applied usually at the macro level.

Tools for statistical matching

70. The list below is based on the CENEX-ISAD WP3 report that offers a detailed discussion of software tools:

- StatMatch (R code)
- SAMWIN
- SPlus code
- SAS code.

C. Other methodological considerations

71. A common issue with linked datasets is inconsistencies in the records linked. Where inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining where an inconsistency arose. It is expected that 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 for the linked datasets.

72. Issues to be addressed by an editing strategy for linked datasets can be summarised by its ability to: edit inconsistencies from the same unit from different sources, treat erroneous and missing variables in a record and ensure consistency in variables across a record for a time period and over time.

73. Sources of potential bias have been identified with regard to integrating datasets. These include:

- Coverage and conceptual issues may only apply for some groups of a population, so care should be taken in generalising results.
- Some variables have the potential to 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.
- The use of linked datasets even for validation purposes may result in a break in the data series that needs to be managed.

74. Extreme care should be taken in backwards and forward casting of 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). A weight may be needed to adjust for missed links in linked datasets.

75. Methods to better estimate linkage errors are required to determine models appropriate to account for these 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 wherein one dataset is sourced from an external dataset since the unit may be defined differently in the external dataset.

76. 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.

Annex 1: Types of data integration

A. Integrating survey and administrative sources

77. The subject of integrating administrative data is not new in the statistical world. But the degree and systems of integrating administrative and survey sources vary greatly across countries; some already have extensive experience integrating survey and administrative data resulting in register-based statistical systems, while others are just starting to integrate the data. In some countries administrative data may have existed for some time but not been used.

78. Practical experience shows that countries who have made the most of the data integration have a tradition in register-based statistics. Register based statistics refers to a system that is based on administrative data and in which statistical registers have been organized into a linked statistical system. A register is defined as a systematic collection of unit-level data organized in such a way that updating is possible. A requirement is that each unit in the register can always be uniquely identified.

79. A system of register-based information was developed first in Nordic countries, with other countries following. It is thoroughly explained in the publication Register-based Statistics in the Nordic Countries - Review of Best Practices with Focus on Population and Social Statistics[27]. Preconditions for such a system are a legal base, unified identification systems and cooperation among institutions while registers are the spines of the practical integration of data from multiple sources.

80. Some examples of survey and administrative data integration are: social surveys (e.g. LFS, EU SILC), various registers (e.g. employment, education registers) and there are several examples of administrative data being combined with survey data for producing indicators traditionally collected through censuses (e.g. agricultural censuses and register based population censuses).

81. The data may be integrated using record linking or statistical matching or may use modelling approaches. It may involve pooling or combining information from multiple surveys, including surveys not conducted by the statistical organisations themselves.

82. There are common challenges faced in the 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 statistical organisation wants to use administrative data in a direct way. In the case of surveys carried out with the use of data from administrative sources it is crucial to gather all data.

83. Administrative data are an already existing source and are already being collected for

[27] UNECE (2007). Register-based Statistics in the Nordic Countries - Review of Best Practices with Focus on Population and Social Statistics. New York and Geneva: United Nations. http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

administrative purposes. Sample surveys require the design of sample frame and contact and response from the target population. Use of administrative data can be cost-effective and cheaper than collecting data by questionnaires. As they already exist there is no additional respondent burden on individuals and businesses. Sample surveys are generally more flexible than administrative sources as they are designed to meet a precise purpose. Administrative sources are on the other hand are the result of a legislative system and may have limitations concerning statistical purposes.

84. Administrative sources can provide full coverage of populations. The ability of administrative data to cover whole populations enables the production of local area data to a level of detail not permitted by sample surveys, which is also of advantage in implementing local policies. Administrative sources can also have the ability to produce more frequent statistics. It depends on the nature of data, but in some cases, such as administrative population registers, business registers, farmer registers and social security data, the sources can be updated daily.

85. In the statistical production process, administrative and survey data can be integrated in different ways. Considering usage of administrative data there is a distinction between direct and indirect usage[28].

86. *Direct usage* is when administrative data supplement or replace the sample survey. When administrative data supplement sample surveys there are more possibilities. Often some variables will be based on questions in the sample survey and some variables will be supplemented from the administrative data. A good example of supplementing sample surveys for a set of variables are the LFS and EU SILC[29]. Another way is that administrative data supplement the sample survey for a part of the population. When administrative data replace the sample survey the statistics are based entirely on administrative sources. Some examples are: employment, earnings, education, agricultural statistics, register based population censuses[30], agricultural censuses.

87. *Indirect usage* of administrative data is when they are used for sampling frames, establishing and maintaining statistical registers, data editing and imputations, data validation and estimation (e.g. small area estimation) and weighting.

88. There are a number of challenges in integrating administrative and survey data. Since administrative data are collected for non-statistical purposes, the difference in concepts might lead to coverage problems as well as bias problems. In some cases, such as business statistics, units do not necessarily correspond directly to the definition of the required statistical units. This requires some modelling 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 understanding of the impact of these differences. Sometimes it is possible to influence the administrative definition by co-operating with the responsible authority. Administrative data are

---

28 https://ec.europa.eu/eurostat/cros/system/files/Usage%20of%20Administrative%20Data%20Sources%20for%20Statistical%20Purposes.pdf

29 http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

30 http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

compliant with the laws in the country and in some instances, definitions and concepts are good enough to provide national statistics but cannot be used in the context of international comparisons to other countries, for example in the ESS system.

89. When administrative data are used to supplement surveys, or are integrated with other administrative data it is desirable that the two data sets contain overlapping information. The ideal situation is if data sets contain unique identifiers. If there are no unique identifiers, combinations of other available individual characteristics have to be considered instead, such as name, gender, address, and date and place of birth, to identify identical subjects in both data sets.

90. Another issue is classifications. Classifications used in administrative data may not be the same as classifications in statistical production. In cases of different classifications, the usual approach is to use correspondence tables and conversion tools based on additional variables that may be available for converting into more correct classification codes. 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.

91. In administrative sources, coding is often done by the respondent, while a sample survey may have open questions and coding is often done by experts. Co-operation between the statistical organisation and the administrative authority is an effective way to solve a part of the classification problem. The statistical organisation can provide experience and may be the one responsible for maintaining the classification.

92. Another issue that concerns classification is a decision to use directly translated international classifications or national classifications. It depends on what national data are needed; however, 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.

93. Missing data and errors also need to be considered. Missing data happen 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.

94. Timeliness is one more point in integrating administrative and survey data. Administrative data may not be available in time or may not coincide with the statistical reference period. Sometimes it can be resolved by analysing the impact and if necessary adjusting it using models.

95. Many international statistical offices and statistical organisations have guidelines, directives, standards, recommendations concerning administrative data. Following is an example of guidelines for dealing with administrative data that can be found on Statistics Canada

webpages (summarized to some extent)[31]:

- Maintain a continuing liaison with the provider of administrative records.
- Understand the context under which the administrative organization created the administrative program (e.g. legislation, objectives, and needs).
- Keep in mind that if the information provided to the administrative source can cause gains or losses to individuals or businesses, there may be biases in the information supplied which can lead to unexpected coverage problems and biases.
- Collaborate with the designers of new or redesigned administrative systems.
- Develop an imputation or a weight-adjustment procedure to deal with this nonresponse (unless non-respondents can be followed up and responses obtained). Administrative sources are sometimes outdated. Therefore, as part of the imputation process, give special attention to the identification of active and/or inactive units.
- In the case when a common matching key for both sources is not available and record linkage techniques are used, select the type of linkage methodology (e.g. exact matching or statistical matching) in accordance with the objectives of the statistical program. When the purpose is frame creation and maintenance, or data editing, exact matching should be used. In the case of imputation or weighting, exact matching should be used, but statistical matching can be also sufficient. When the sources are linked for performing some data analyses that are impossible otherwise, consider statistical matching, e.g. matching of records with similar statistical properties.
- When record linkage is to be performed, make appropriate use of existing software.
- When data from more than one administrative source are combined, pay additional attention to reconcile potential differences in their concepts, definitions, reference dates, coverage, and the data quality standards applied at each data source.
- Some administrative data are longitudinal in nature (e.g. income tax, goods and services tax). When records from different reference periods are linked, they are very rich data mines for researchers. Remain especially vigilant when creating such longitudinal and person-oriented databases, as their use raises very serious privacy concerns.
- Use identifiers with care, as a unit may change identifiers over time. Track down such changes to ensure proper temporal data analysis. In some instances, the same unit may have two or more identifiers for the same reference period, thus introducing duplication in the administrative file. If this occurs, develop a mechanism to remove duplicates.
- Document the nature and quality of the administrative data once assessed. Documentation helps statisticians decide the uses for which the administrative data are best suited. Choose appropriate methodologies based on administrative data and inform users of the methodology and data quality.

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

96. In recent years, the official statistics community has acknowledged the value of big data and has been exploring the use of diverse sources in several domains. Many different types of data sources fit under the umbrella of big data. One example is scanner data on prices, coming

[31] Statistics Canada. Use of Administrative data. http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm#a2

from scanner transactions in supermarkets and often provided to statistical organisations by private companies working in marketing. Another example is data scraped from the internet. Official statistics has acknowledged the value of Internet-scraped data and has been exploring their use in several domains (for instance in statistics on ICT use in enterprises and tourism). Data can be scraped directly from individual websites, but this approach requires first identifying the websites and then dealing with different queries and different formats obtained from each website. Alternatively, data can be scraped from "hub" websites describing a plurality of units (for instance hotels data), although the information available may be summarised.

97. It is quite common to state that big data provides information useful for statistical purposes in a way that is substantially cheaper, faster, more timely than survey and administrative data. However, it is not always recognised that the relevance of data coming from the new sources should be investigated first. Moreover, the introduction of big data approaches, i.e. data provider agreements, new IT tools and capabilities, can also be very expensive and time consuming, jeopardizing at a first step the advantages of the big data usages.

98. One significant opportunity arises from the global nature of some big data: the opportunity for statistical organisations to collaborate on crafting global data agreements and global partnerships with big data providers.

99. Consequently, the constraints/limitations in the usage of big data should be properly understood, at the design stage of the data integration activities. As usual, innovation requires acceptance of some risks, but those should be clearly understood, stated and managed to mitigate them.

100. A way to mitigate risk could be to focus on expectations, making them as clear and reasonable as possible. Moreover, the use of big data requires flexibility agile approaches, due to often unexpected changes in the source data. For example, in web scraping website changes and data layout changes can occur without warning. Good relationships and agreements with data providers may help in managing these situations; however, it is important to consider in advance what might go wrong and how to react.

101. For some of these external sources, the reported objects can be easily associated with statistical units of the target population. On the contrary, there are cases in which the big data objects need to be elaborated in order to be compared to statistical units. In most cases, when big data are not directly comparable with data collected and organized by statistical organisations, a lot of work is needed to create integrated data. Finally, sometimes the big data offers information on topics that are not well covered by traditional surveys - in this case the advantages in their use is unquestionable.

102. In addition, it is likely that the big data are not standardized or codified. In some sense, official statistics are already prepared to face this kind of problem, but the pre-processing phase is a very time-consuming process and a lot of work is needed to identify models that can easily support data reconciliation, management of the complexity and to allow the data integration step. In order to integrate big data in the statistical production process, a system is required for data ingestion and reconciliation that allows managing a big data volume of data coming from a variety of sources. The statistical production system needs to produce the ontology and the big data architecture, and the mechanisms for the data verification, reconciliation and validation.

103. In the cases of coincidence or harmonization between big data objects and statistical units, if a unit identifier is available and shared with the statistical organisation, the big data can be integrated with existing statistical data at micro-level, so to enlarge the content, the coverage, the accuracy and the timeliness of official statistics. When identifiers are not available, big data can be used in combination with other sources at aggregated levels.

104. When using and integrating new data sources, new methods may need to be developed and integrated with the existing ones. The opportunity to study and develop new methods requires some patience to allow them to evolve and to become stable. In this spirit, it is important to not leave research works in the drawer when they don't produce positive or expected outcomes, so that other groups don't replicate unsuccessful experiences.

105. It is important to collaborate with developers/administrators of sandboxes and big data technologies. The IT sector is strongly involved in the modernization program looking at the use of big data. Due to their characteristics, big data are often difficult to integrate into existing systems so costly changes to IT infrastructures may be necessary.

106. The usage and integration of new data sources require a composite team of skills and professionalisms. The best would be a team composed by experts from methodology, IT, social-media, subject matter, new tools (e.g. web scraping, visualization)

107. As for administrative data, it is important to assess the quality of the input, the throughput and the output, however, sometimes the input it is not under full control of the statistical organisations as well as the procedures in the data processing steps are not fully understandable by traditional skills. In these cases, a good relationship with the data providers are important to understand the data, so as definitions and concepts behind the data will help in evaluating their quality.

108. A framework for quality assessment when using big data seems to be produced in compliance with already existing quality framework, e.g. in sample survey or in combined use of administrative data. As far as the quality dimensions, it is often noticed that the big data may suffer of coverage issue, being not representative of the target statistical population

109. When using and integrating big data in the statistical production system, it is necessary to proceed in steps, starting with clearly stating the focus of the analysis in which the big data are involved. Consultation with experts around the world, contacts with others approaching the same tasks, learning lessons from other experiences are important points in order to design further steps, as well as understanding risks, and fix what is necessary in terms of skills, IT capabilities and methods.

110. The good relationship and agreement with data providers needs to be established, when the data are not directly collected by the statistical organisations. Then an experimental stage should follow, in which a sample test data should be requested, as well as several new skills, IT tools and methods should be experimented in a common setting, together with well-established tools and methods. The experimental step will require flexibility and patience in evaluating the potentialities of the outcome. Some efforts should be devoted to the quality evaluation and to share also intermediate results with others.

111. At the end, after setting some conditions on how much research should be done (e.g. what kind of quality is acceptable), there will be possible to introduce the new data, tools and methods into the production system, also via comparison with existing results coming from traditional data sources.

C. Integrating geospatial and statistical information

112. Statistical data is almost always related to a certain physical space, like a municipality, a state, a country a region, etc. 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, communications infrastructure, facilities are examples of geographic information which are indispensable elements to fully understand the figures that official statistics generates.

113. The geospatial and statistical data integration landscape is complex. The Global Statistical Geospatial Framework (GSGF - UN GGIM) and initiatives such as GEOSTAT projects (Eurostat) are vital for developing a consistent and systematic approach to linking geospatial and statistical data.

114. The Global Statistical Geospatial Framework (GSGF)[32] is a high-level framework which facilitates consistent production and integration approaches for geo-statistical information. It is generic and permits application of the framework principles to the local circumstance of individual countries.

115. An example of good practice of assessing the maturity and capability of organizations for spatial statistics is the decision tree developed in the GEOSTAT 2 project. It is a path of practical dynamic questions to be answered before embarking on integrating geospatial data with statistical data as shown in the figure below.

[32] http://ggim.un.org/ggim_20171012/docs/meetings/GGIM6/Background-Paper-Proposal-for-a-global-statistical-geospatial-framework.pdf

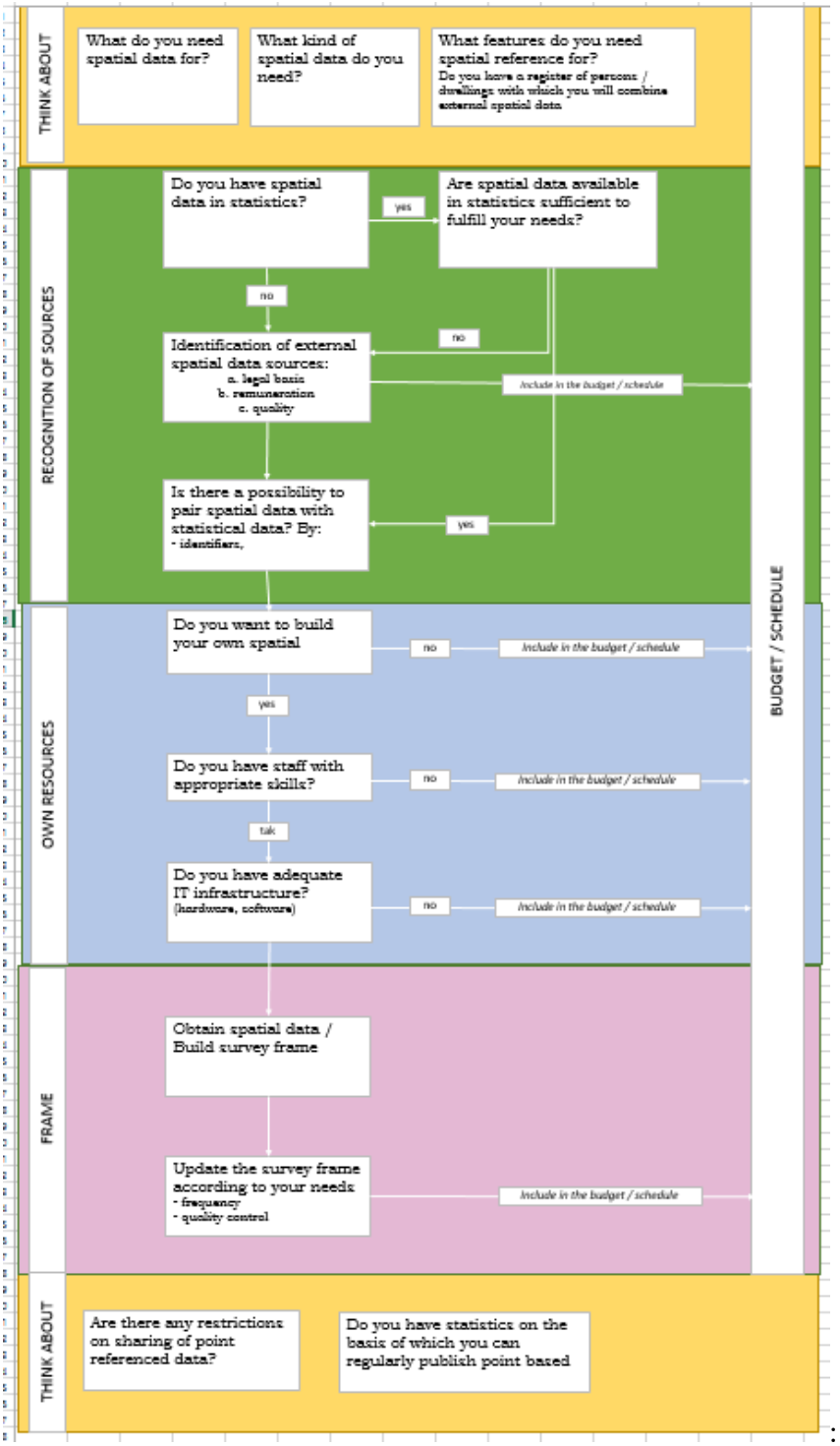

Figure 3. Decisions related to integrating geospatial data

116. Developing a coherent and systematic approach to linking statistical and geospatial data is likely to take some time. The best way to achieve consistent integration is through having a common method of geospatially enabling statistical and administrative data and integrating this with geospatial information through an internationally agreed the Global Statistical Geospatial Framework, which enables comparisons within and between countries.

117. Integration can take place at any stage of the statistical production process, as described by the GSBPM. The integration includes geocoding of statistics, spatial analysis, and creating statistical maps. As part of the integration process the following steps may occur:

- Geocoding statistical information at unit-record level
- Processing and manipulation of statistical information using spatial analysis techniques with the purpose of selecting information or deriving new information with a focus on their spatial characteristics, e.g. buffering around spatial features
- Supporting a more efficient and flexible statistical production process with geospatial information e.g. for surveying and sampling, field operation
- Combining statistical end products with geospatial information in statistical maps
- Improving the quality of existing statistical products adopting spatial models, e.g. commuting information by calculating journey times based on detailed transport networks.

118. All statistical phenomena that can be associated to a location are in principle relevant for the integration of statistical and geospatial information. Location in this context means the location of the most individual observation at unit record level. In most cases the location will be a point with coordinates or a precise address. However, other spatial reference frameworks like lines or polygons are relevant as well representing e.g. road segments or areas with a certain land cover.

119. Integration of geographical data with statistical data aims to improve the value of the statistical information that is being produced. Geographic information systems (GIS) as far as it is possible should be used at all stages (inventory, preparation, progress, monitoring, dissemination of results) of the geospatial integration. Wherever it is possible, data should be collected with reference to an address point - results can then be disseminated using any desired spatial divisions. GIS technology should be considered only at a level appropriate to the skills and resources available and constitute an integral part of the overall work of a national statistical organization.

"The 10 Level Model"

120. An example of a detailed practical model is "The 10 Level Model" developed by CSO Poland. It can be used to better understand and develop a statistical and geodetic reference framework as a standard of geospatial data production.

121. Recently Statistics Poland worked on the project which aim was the improvement of the use of administrative sources. As a result of the Polish experience, the methodology of assessing the usability of administrative data sources has been elaborated. Nevertheless, quality assessment should be conducted separately for each register, taking into account its possible use. The methodology of assessing register quality will include a few sections. For purpose of statistical division based on geodetic division the section regarding information about the quality of spatial data register will be the most important. The issues included under this section will enable assessing the overall quality of data sets, and the quality of data which they include. Within this area two criteria have been distinguished, i.e. accuracy and comparability which will be measured by specific indicators.

122. Polish methodology of assessing spatial data sets could be a proposal of standard for other countries and statistical organisations which want to harmonise statistical and geodetic divisions to receive better quality of statistical geospatial data and analysis.

123. The opportunities arising from integration of geospatial data with statistical information include:

- integrating the sets of data from different sources of data ex. administrative data educational data, mobile data with geospatial information
- increased added value of statistical and spatial data
- better quality of geospatial data, integrated statistical and spatial data
- better interoperability of sets of data, possibility of analyses, easier methods for linked data sets
- better quality and different spatial analyses
- many cartographic methods of presenting data
- new kind of services and data for users' needs
- flexibility to use statistical information by external users
- useful for policy and decision makers, especially for regional policy makers
- useful for scientific purposes
- useful in environmental protection
- enhanced collaboration between mapping agencies with statistical institutions with maintenance timeliness sets of data and systems.

124. The challenges include:

- sources of data, quality of data
- format files and reference system of the data sets from different sources
- differences in classification of territorial units among countries outside EU
- the reference time of the data
- aggregation level of the data sets
- budget restrictions
- legal issues
- expertise and knowledge
- confidentiality
- accessing skills in GIS and other geospatial related areas
- the need for additional technology to prepare geospatial data and to publish results
- standardization of identifiers (or other information by which it is possible to link statistical information to geospatial information)
- collaboration between mapping agencies with statistical institutions and other institutions (scientists)
- testing/exploration of data source before getting further details.
- secure ways to exchange the data sets
- secure ways for processing data sets
- close cooperation is needed with geospatial data providers (administration, mainly National Mapping Agencies, other organisations).

D. Validating official statistics

125. 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[33]. 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 statistical organisations.

126. It is possible to use external data sources to determine accuracy of survey results or use survey results to challenge results from alternative data sources of providers of statistics either at the micro level, i.e., linking information from multiple data sources on an individual person or business firm (unit) or at a macro level, i.e., linking information from multiple data sources on a group of people or business firms (units).

127. The issues involved in integrating alternate data sources into the validation processes used for producing official statistics include:

- assessing origin and quality of the source, including 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).

128. There are a number of related initiatives:

- 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[34].
- ESSnet project on Data Integration. This completed project focused on the methodologies and methodological issues of micro data integration [35].
- ESSnet project Integration of Survey and Administrative Data. The project aimed at developing the knowledge and expertise of participating statistical organisations in the use of integrated survey and administrative Data in the production of official statistics[36].
- ESSnet project on macro-integration. This project discusses various methods of integrating data sources at aggregated or macro level[37].

129. The following are paragraphs outline the opportunities that data integration has provided at Statistics New Zealand to validate statistics.

---

[33] http://www.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf
[34] https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en
[35] http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en.
[36] http://ec.europa.eu/eurostat/cros/content/isad-finished_en.
[37] https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

130. The advancement of data integration skills has also led to the creation of Statistics NZ's Integrated Data Infrastructure (IDI). The IDI brings 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. The IDI has paved the way to answer complex research questions to improve outcomes for New Zealanders.

131. Administrative data have been linked to examine and decide on their specific use in the production of official statistics. Inland Revenue data, specifically longitudinal payroll data from the Employer Monthly Schedule (EMS) returns was linked to produce new statistics - filled jobs, worker flows, and total earnings - that measure labour market dynamics at various levels – including industry, region, territorial authority, business size, sector, sex, and age. These statistics provide an insight into the operation of New Zealand's labour market.

132. Data integration has also been used for the improvement of a survey process as illustrated in the linking of the March 2013 Household Labour Force Survey (HLFS) to the 2013 Census data to analyse non-respondents to the HLFS. The project led to the deletion of a non-response adjustment step in the weighting procedure for the HLFS simplifying the HLFS estimation process.

133. Some validation projects involving the use of various administrative data sources have led to recommendations of using these data sources for either benchmarking income survey results, imputation or validation of income statistics rather than using the administrative data sources to replace various sources of income. The administrative data sources need not be integrated to the income surveys when using them for benchmarking or validating income statistics. In cases where data integration will be required for the above immediate uses, a new process – data integration – will need to be designed in the production process.

134. Linking the Census to administrative data sources in the IDI has been instrumental in the realisation of some of the goals of Statistics NZ's Census Transformation Programme. The programme is investigating alternative ways of running New Zealand's future census including the feasibility of using linked administrative data to replace census questions.

135. Data integration has also paved the way in the development of new methods, e.g., new models. One good example is the production of population estimates using administrative data. Bryant and Graham (2015) use Bayesian modelling to estimate, specifically, regional populations in New Zealand based on administrative data on birth and death registrations, tax and NZ international passenger movements.

136. Integrating multiple data sources face a number of challenges. A number of these are described in the following paragraphs.

137. Timeliness of external data sources, unless receipt of data is common and regular, will always be a challenge for linked datasets and for statistics produced from these datasets. These include:

- The promptness in picking up birthed units to an administrative data source. A high number of birthed units not picked up quickly enough by administrative data leads to potential bias in official statistics.

- Timing issues around getting the linking accomplished in time for production.

138. The cooperation of the dataset owner is also another challenge to address. The statistical organisation needs to ensure the continuity and consistency of the quality of the data to be provided. However, contingency plans need to be in place in case the data source becomes unavailable. The statistical organisation may also need to elicit assistance in determining the definition of concepts, classifications or populations in case these need to be redefined to better suit their needs.

139. After quality assessment of an external data source has been undertaken, the next challenge to address is the extent an external data source will be used to meet the statistical need. Are new methods required to convert the external data source into a form useful in the production of a statistical output?

140. Although administrative data may be freely available to a statistical organisation, other external data sources may not necessarily be available for free. Costs may also be a challenge in accessing external data sources. Costs are also incurred in the quality assessments of external data sources and all these costs need to be determined and assessed before proceeding with any data integration project.

141. Another challenge is the resistance to changing any part of a 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.

142. The need for standardised processes which are responsive to administrative changes in the data supplied and to new administrative data available to Statistics NZ should also be addressed when using external data to validate official statistics.